The Sense of Agency: Underlying Neurocognitive Mechanisms and its Attribution to Human and Non-Human Co-Actors

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

im Fach Psychologie

Eingereicht an der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

M.A. Michael Goldberg

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Bernhard Grimm

Gutachter/Gutachterin:

Prof. Dr. Elke van der Meer
Prof. Dr. Niko Busch
Prof. Dr. Dorit Wenke

Eingereicht am 29.09.2017
Verteidigung am 21.02.2018

Content

Acknowledgement

The work presented here would have not been possible without the support from a professional and personal network. With this acknowledgement I would like to express my sincere gratitude to everyone involved.

I would like to thank my advisors, Prof. Dr. Elke van der Meer and Prof. Dr. Niko Busch for guiding me throughout the whole process. Thank you for the scientific advice, for sharing from your knowledge and time and for supporting where difficulties or uncertainty arose. Your kindness and optimistic outlook meant a lot and contributed to the project's success.

In addition, I would like to thank Dr. Gesa Schaadt for teaching and introducing me to the complex world of EEG and for a general guidance as a doctoral candidate. Thank you for your endless will to help, for sharing your knowledge, for listening and supporting, for giving tips and helping solving problems along the road. Your advice, criticism and sharp mindset helped me a lot to walk more confidently in places I was just discovering.

I would also like to thank Christina Rügen, the technical assistant of the EEG lab, for being so kind, patient and thorough during the stage of my data acquisition. Thank you for helping me to communicate with participants and overcome the language barrier at very early stages. The same goes to Michelle Wyrobnik for offering help at many different points of the project.

Finally, I would especially like to thank my close colleague and friend Florian Koller for creating a very supportive and friendly working environment. I thank you for choosing to cooperate with me on two projects and for having so much patience and trust. Your scientific knowledge, analytical skills, and warm character enriched and pushed our work forward. Thank you for being there for me at all times both professionally and personally.

Academic Curriculum Vitae

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne unerlaubte Hilfe verfasst habe;

dass ich die Doktorarbeit an keiner anderen Universität eingereicht habe und keinen Doktorgrad in dem Promotionsfach Psychologie besitze;

und dass mir die zugrunde liegende Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II vom 17.01.2005, zuletzt geändert am 13.02.2006, veröffentlicht im Amtlichen Mitteilungsblatt der HU Nr. 34/2006, bekannt ist.

Michael Goldberg
Berlin, den 29. September 2017

Zusammenfassung

Das Gefühl der Kontrolle über die eigenen körperlichen Handlungen, und dadurch über die externe Umwelt ist einer der Grundpfeiler unserer menschlichen Existenz. Dieser fundamentale Aspekt der Identität ist bekannt als 'Sense of Agency' (SoA). Innerhalb der Neurowissenschaften begann die intensive Untersuchung dieses faszinierenden Konzepts erst innerhalb der letzten zwei Jahrzehnte. Das vorliegende Forschungsprojekt befasst sich mit zwei zentralen Aspekten des Sense of Agency. Zum einen wurden die zwei zugrundeliegenden neurokognitiven Mechanismen 'Vorhersage' und 'Retrospektive Inferenz' untersucht. Dazu wurden die Bewältigung einer Zeitschätzaufgabe zu selbst ausgelösten Handlungen und Handlungsfolgen basierend auf dem klassischen Libet-Paradigma als implizite Messung für Sense of Agency mittels des Intentional Binding Effekts und elektrophysiologische Messungen mittels eines EEGs kombiniert.

Die Ergebnisse zeigten, dass beide Mechanismen eher durch den unmittelbaren als den längerfristigen Handlungskontext, im Sinne der Anzahl der vorausgehenden Handlung-Folge-Paarungen, erklärt werden. Zudem zeigten sich Modulationen sowohl des Bereitschaftspotentials als auch des auditorisch evozierten Potentials und somit reflektierte neuronale Dynamik, die mit den beiden Mechanismen einhergeht.

Zum anderen wurde die Zuschreibung von Agency bei weiteren Ko-Akteuren, mit denen eine gemeinsame Aufgabe bewältigt werden musste untersucht. Innerhalb eines ökologisch valideren Kontexts, einer Fahrsimulation, die eine reale Fahrsituation abbildet, konnte diese Zuschreibung erst bei menschlichen und nicht-menschlichen Ko-Akteuren, außerdem bei einem vermenschlichten Computer und einem vor der Aufgabe selbst trainierten Computer, verglichen werden. Es zeigte sich (erwartungsgemäß), dass Agency auf die Handlungen menschlicher jedoch nicht auf die Handlungen nicht-menschlicher Ko-Akteure erweitert wurde. In Bezug auf die Handlungsfolgen wurde Agency bei beiden Ko-Akteuren erweitert. Interessanterweise zeigte sich zudem, dass Agency im Falle des vorher selbst trainierten Computers als nicht-menschlichem Ko-Akteur in derselben Weise wie bei einem menschlichen Ko-Akteur zugeschrieben wurde. Wohingegen das für die Zuschreibung von Agency bei der gemeinsamen Aufgabenbewältigung mit dem vermenschlichten Computer nicht galt.

Das durchgeführte Forschungsprojekt trägt somit zu einem tieferen Verständnis menschlicher Agency auf individueller Ebene und im sozialen Kontext bei. Außerdem liefert es Implikationen für die Mensch-Maschine-Interaktion und die Verbesserung zukünftiger Mensch-Maschine-Schnittstellen.

Summary

The seamless feeling of control over one's own bodily actions, and through them, over the external environment is one of the cornerstones of our existence as human beings. This fundamental aspect of personal identity has been termed the sense of agency (SoA). It is only within the last two decades that this intriguing concept has begun to be intensively studied in the cognitive neurosciences. In the current research project we addressed two central aspects of the sense of agency. First, we investigated its underlying neurocognitive mechanisms: prediction and retrospective inference. To that purpose we combined a behavioral task based on the classic Libet paradigm with electrophysiological recordings (EEG). Temporal estimations of self-produced actions and outcomes were collected to calculate the intentional binding effect, an implicit measure of the sense of agency.

Our results suggest that the immediate rather than the long-term context of an action, in terms of the amount of preceding action-outcome couplings, can account for both mechanisms. Furthermore, we have found modulations in both the readiness potential and the auditory evoked potential that reflect the neural dynamics associated with these mechanisms.

Second, we looked into the attribution of agency to other co-actors when cooperating in a joint task. Using a driving simulator, we created a new ecologically valid environment and compared agency attribution to human and non-human co-actors. Moreover, we tested two types of non-human co-actors: an anthropomorphized computer, and a computer that had been trained prior to the joint task. We found that agency is extended over actions of the human partner, but not over those of the computer. In contrast, outcomes were not affected by the type of co-actor and were extended in both cases. Interestingly, a computer that had been trained by participants prior to the joint task was attributed with agency in the same way as a human partner. However, this was not the case with an anthropomorphized computer, which showed to be no different to a normal computer.

Overall, the current research project has made a step towards a better and deeper understanding of human agency in the individual as well as the social contexts. Additionally, the findings presented in this work inform the field of human-computer-interaction and contribute to the improvement of future interface designs.

List of Original Research Articles

Study I:

Goldberg, M., Busch, N., and van der Meer, E. (2017). The amount of recent action-outcome coupling modulates the mechanisms of the intentional binding effect: A behavioral and ERP study. *Consciousness and Cognition*, in press.
doi: 10.1016/j.concog.2017.07.001

Study II:

Goldberg, M., Koller, F., Busch, N., and van der Meer, E. (Submitted). The Attribution of Agency to Non-Human Co-Actors in a Joint Task: A Driving Scenario. Submitted to *Cognition*.

Study III:

Goldberg, M., Koller, F., Busch, N., and van der Meer, E. (Submitted). Attribution of Agency to Automated Entities: Humanized versus Trained Systems. Submitted to *Cognition*.

List of Abbreviations

SoA	Sense of Agency
IBE	Intentional Binding Effect
ToM	Theory of Mind
EEG	Electroencephalography
ERP	Event-Related Potential
RP	Readiness Potential
AEP	Auditory Evoked Potential
ANOVA	Analysis of Variance
ICA	Independent Component Analysis

List of Figures

“Let us not forget this: when 'I raise my arm', my arm goes up. And the problem arises: what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?”

Ludwig Wittgenstein
(Philosophical Investigations, 1953, §621)

1 Introduction

1.1 The Sense of Agency

The sense of agency (SoA) refers to the feeling of control over ones own bodily actions, and through them, the outcomes in the external environment. The term emerges within the broader philosophical and psychological debate of selfhood and personal identity. Throughout the years, a large number of models, types, aspects and delineations have been offered with regard to the concept of the 'self'. To mention one famous example, we can think of William James' categorization of the physical self, mental self, spiritual self, and the ego (James, 1890). Another more recent dominant example is the distinction of the abstract symbolic self (Sedikides & Skowronski, 1997; Kihlstrom et al., 2003) that stands in contrast to the embodied social self (Niedenthal et al., 2005; Barsalou, 2008). While being important milestones in the theoretical and empirical research fields of the self, it is beyond the scope of the current work to consider all these different conceptualizations. Instead, I would like to focus on a narrower distinction that serves as the underlying theoretical ground for the concept of agency as it will be discussed in the following sections of this work.

Drawing on the interface between philosophy of mind and the cognitive sciences, Shaun Gallagher presented in a review paper from 2000 a distinction between the minimal and the narrative self. The minimal self is defined as the most basic, immediate, primitive sense of the self. This self is the immediate subject of experience and is unextended in time. Furthermore, the minimal self is not informed by conceptual thought and does not need to know or be aware of its experience in order to maintain its identity. It is therefore, the self as it is given to itself without any mediation or interpretation. In contrast, the narrative self is extended in time and is created by building a coherent self-image. This extended self is constituted on memories and stories of the past as well as intentions and hopes for the future. Hence, it is more closely related to what we intuitively refer to when we talk about ourselves as "I ", "me" or "myself".

The sense of agency is, according to Gallagher (2000), one aspect of the minimal self. As defined above, the minimal self is first and foremost an experiencing self. Gallagher further distinguished between two closely related aspects of experience:

On the one hand, the sense that I am (i.e., my body) undergoing an experience and on the other hand, the sense that I am intentionally acting in the experience. Gallagher names the former a 'sense of ownership' over ones own body, and the latter a 'sense of agency', that is, the feeling of being the initiator or source of the action. At first glance, it might seem counter-intuitive to distinguish between the two. Indeed, in most daily experiences of voluntary or willed action, the two senses coincide and cannot be separated. For example, consider the action of moving ones hand forward reaching for a bottle of water that is placed on the table. In this case, it is me who both initiated the action (agency) and the one that undergoes the movement (ownership). However, in cases of unwilled or involuntary action, the two come apart. Consider now being pushed by someone else, or the case in which a physician is manipulating a body part in a medical examination. In these cases it is easier to see how a sense of ownership is retained (i.e., I know that *I* am, or *my* body part is being moved) while no sense of agency is formed (i.e., I have no feeling of causing or controlling the movement). In a more recent paper, de Haan and de Bruin (2010) suggested to reconsider this distinction between ownership and agency. The two researchers challenge the idea that ownership and agency are mutually exclusive theoretical aspects of the minimal self and argue for a gradual difference that always includes both aspects to some extent. While this debate is still open-ended, and keeping in mind conceptual reservations, we shall focus in this work on the SoA and consider it as a distinct aspect of selfhood, putting the emphasis on the experience of volition and intentionality.

Whereas selfhood and personal identity have been in the center of elaborated philosophical and psychological discussions for many centuries (among others see Locke, 1694; Shoemaker, 1984; Nichols & Bruno, 2010), the SoA has only recently started to be intensively researched. One central reason for this lag is the lack of any experimental paradigm or operational measurement that would have enabled the scientific study of the theoretical concept. A seminal work by Benjamin Libet (1983) opened this door, and became an important landmark in the study of volitional action and the relation between intention and action. Libet and colleagues showed for the first time, that although intentions were reported to occur prior to actions, brain activity could already be noticed preceding subjects' awareness of their decision to move. Libet's findings opened up an extensive discussion in both philosophy and cognitive science about free will, the lack of it, and its implications on decision-making, responsibility, and moral action (see Searle, 2001 for a philosophical discussion and

Klein, 2002 for a neurocognitive commentary). More relevant to our work is rather the Libet-clock paradigm that two decades later led to the development of the measurement for the SoA. In Libet's study, participants watched a rotating clock and judged the time of different spontaneous events (e.g., the time in which they moved their hand, in which they became aware of the intention to move it, etc.). The reported times were compared to the clock times and judgment errors were calculated. Brain activity was recorded simultaneously using electroencephalography (EEG).

It is this classical paradigm (the Libet-clock) that was employed in 2002 by Haggard and colleagues to study the perceived times of actions and their sensory outcomes (Haggard et al., 2002a, 2002b). Haggard and colleagues designed agency and non-agency conditions: on baseline non-agency conditions participants either pressed a button (action) or listened to a tone (outcome). Action did not result in an outcome and outcome was independent of action. In operant agency conditions participants pressed a button that was always followed by a tone (i.e., action followed by outcome). Within each trial, participants watched the rotating clock and were asked to estimate the perceived time of either the button press or the tone. Judgment errors (i.e., distance between perceived time and clock time) were compared between baseline non-agency and operant agency conditions. The central finding of the study was that in operant agency conditions, action times were shifted forward and tone times were pulled backward compared to baseline non-agency conditions. In other words, the derived time interval between action and outcome on operant conditions was compressed in comparison to the same interval on baseline conditions. This was the first evidence linking between modulation in time perception and agency. Haggard and colleagues (2002b) were also interested to find out whether the temporal effect is unique to voluntary actions and created a third condition to test this question. Involuntary movements (finger twitch that caused a button press) were induced via transcranial magnetic stimulation (TMS) over the primary motor cortex of participants. Once again the same comparisons were made and results showed that no temporal compression was found when actions were involuntarily induced. Actually, the binding of action and effect was reversed and the time interval increased in operant compared to baseline conditions. Taken together, this seminal work has reported for the first time an implicit measure for the SoA, termed the intentional binding effect (IBE). Since then, the effect has become a key player in the study of the SoA, employed in an ample amount of studies in the field of human agency (see a recent review by Haggard et al., 2017).

1.2 The Underlying Mechanisms

While the SoA has been extensively studied over almost two decades, its underlying neurocognitive mechanisms are still not fully understood. However, by looking at the literature so far, two central positions should be differentiated with regard to the origins of the SoA: prediction and retrospective inference. It is still a matter of debate to what extent each of the two contribute to the emergence of agency and in what way the two mechanisms relate to one another (for a theoretical framework that goes beyond this dichotomy see Moore et al., 2009a). Let us elaborate on each of the two mechanisms.

According to the predictive approach, processes dedicated to the control and preparation of voluntary action determine the sense of agency. Predictions of both the future states of the motor system as well as the sensory outcomes of a movement are required for the execution of a seamless motor behavior and learning. These processes are termed internal forward models and can be divided into two kinds: forward dynamic and forward sensory models (e.g., Blakemore & Firth, 2003). As could be expected from their names, the forward dynamic model is responsible for the monitoring and correction of bodily movements on the go, while forward sensory model captures the causal relations between bodily movements and sensory outcomes. It is the forward sensory model that generates predictions of the expected sensory outcome of a specific movement based on the creation of efference copies of the motor command. The comparator model of the sense of agency (see Figure 1) details the way in which the dynamic and sensory models work. Crucial is the comparison made between desired and actual outcomes (the dynamic model serves as a preparatory stage for this comparison). When a match between expected and actual sensory outcomes occurs, a SoA is generated. When a mismatch is experienced, the source of the outcome is not attributed to oneself and no SoA is generated (David et al., 2008).

The second approach to explain the emergence of a SoA is that of a retrospective inferential process. In contrast to the predictive mechanism, the emphasis here is moved from the motor system to the perceived sensory outcome. By relying on sensory information, the brain retroactively infers the causal origins of actions and their outcomes. A central theory that supports this view is the theory of apparent mental causation formulated by Wegner and Wheatley (1999).

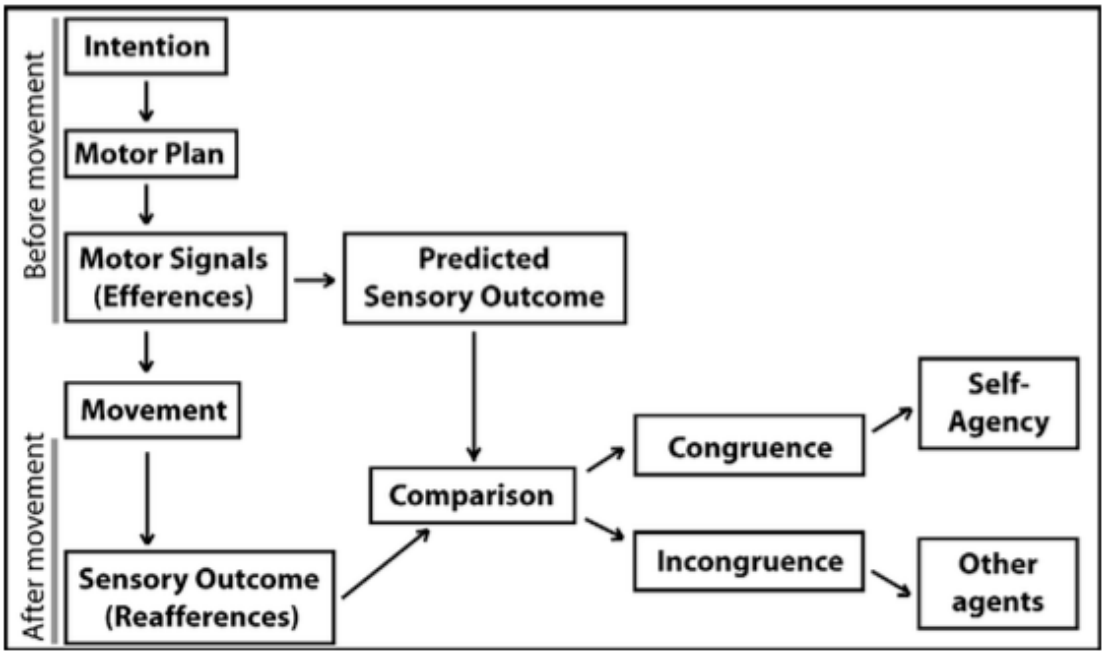

Figure 1: The comparator model The figure shows a schematic diagram of the formation of a SoA based on internal forward model. The sensory model uses an efference copy, that is, a copy of a motor command, to predict the respective sensory outcome of the movement. Congruence of the predicted outcome with the reafference signals lead to the attribution of the SoA to oneself. Incongruence indicates another agent as the cause of an action and outcome. (Adapted from David et al., 2008)

According to the theory's central premise, an intention to act is experienced by the agent only if three conditions are met. The intention has to occur before and be in proximity to the action (priority), the intention has to be compatible with the executed action (consistency), and the intention has to be the most plausible cause of the action (exclusivity). The sense-making process will lead to self-attribution of agency only if these prerequisites are fulfilled. Empirical support for a retrospective mechanism comes from studies showing that explicit judgments of agency are readily biased. In one such study, participants were primed with the outcome of a subsequent action below the threshold of conscious perception. As a result, participants showed stronger SoA over the performed actions (Aarts et al., 2005). In another study, action outcomes were primed using words consistent with the outcome. The researchers found that priming increased a feeling of control over actions that were not performed by participants (Wegner & Wheatley, 1999). Such findings suggest that an inferential process takes place after the action-outcome pairing has been experienced and that computing a SoA relies less on internal motor signals and more on external cues.

In study I, we were interested in extending the research about prediction and inference by uncovering the neural underpinnings of these two mechanisms. Furthermore, we wanted to address the question of how the immediate compared to the long-term context of action-outcome pairings modulate the SoA through both mechanisms. To that purpose, a behavioral task and EEG recordings were combined and the Libet-clock paradigm was employed.

1.3 From a Single Actor to Joint Action

To this day, most studies about the SoA focused on the single actor. Normally, a participant would be seated in the lab and asked to perform a specific task that included an action followed by an outcome. Depending on the paradigm, the participant would then have to give estimations of the extent to which she felt control/agency over the performed action and outcome (either implicitly or explicitly). It is only in recent years that the research of the SoA has been extended to the context of a joint task, including more than a single actor. Joint task experiments have existed long before the inquiry of the SoA, but it is the direct measurement of agency when acting in cooperation with another co-actor that enabled this new line of research.

Experiments conducted in this regard have generated a variety of interesting insights relevant for co-action (see among others Wohlschläger et al., 2003; Tsai et al., 2006; Dolk et al., 2014). When it comes to the SoA, findings seem to point to an intriguing phenomenon: when two people act together in a joint task, a new agentic identity is formed. This new agentic identity, a ‘we’ rather than an ‘I’, can be interpreted as some sort of unified agency that emerges and extends beyond the single actor (Obhi and Hall, 2011a). First direct evidence of this phenomenon was demonstrated by Strother and colleagues in 2010. In their study, two participants were seated one next to the other and asked to press a button in an alternating fashion, such that on each trial only one of them acted. The participants watched a rotating clock and had to estimate the times of actions (either self produced or of the other participant) and their outcomes (tones). By measuring the intentional binding effect, Strother and colleagues found that participants experienced agency not only over their self produced actions and outcomes but also over those of the co-actor. Interestingly, explicit judgments of agency diverged from the implicit, pre-reflective agency as measured by the IBE.

In a later experiment by Obhi and Hall (2011a), an identical paradigm was employed, the difference being that in this experiment both participants had to act. In one of the conditions, participants were instructed to initiate an action within a given time frame. In case that one of the co-actors pressed the button first, the other participant was required to respond and press it immediately after. The time of the first of the two actions had to be estimated. Obhi and Hall found that regardless of being in the role of the initiator or the responder, IBE was present for both actions and outcomes. Again, subjective reports of agency did not extend in the same fashion as the implicit measurement. In another condition, participants were assigned in advance to the role of either the initiator or the responder. The same results were found as in the co-intention condition. Taken together, these experiments showed for the first time that a partial participation in a joint task, having an intention to act or simply being prompted to act by an external event are sufficient for the formation of an extended pre-reflective SoA in the context of a joint task.

In study II we were interested in testing the attribution of agency to another co-actor in a more naturalized ecologically valid environment (compared to the neutral setting employed by Obhi and Hall). Additionally, we wanted to compare between a human and a non-human co-actor. We therefore created a realistic driving scenario and used a driving simulator where participants performed in a joint task with either a confederate or with the computer (autonomous car).

1.4 The Attribution of Agency to Non-Human Co-Actors

Another important recent advancement made in the experimental research of human agency is its inquiry within the context of human-computer interaction (HCI). Although the SoA has been extensively studied in relation to many different psychological factors (e.g., reward, emotion, psychopathology etc.), only a paucity of studies address the experience of agency with regard to automated systems (see Limerick et al., 2014 for a comprehensive review of the topic). Since interaction with computers, machines and robots is now an inseparable part of everyday life, this new exciting field of research is becoming all the more relevant. So far, only a few aspects of the interaction between human and non-human agents have been addressed. For example, researchers have compared two different input modalities (normal button

press versus skin-based input device) and measured its influence on our sense of control (Coyle et al., 2012). A system's feedback is another factor that has been tested. Farrer and colleagues (2008) created distortions in a computers visual feedback that resulted in misattributions of agency with regard to the source of an action.

Through the combination of the newly designed paradigms for multiple co-actors in a joint task together with the emerging branch of HCI studies, a more direct comparison between different types of co-actors has become available. In a follow-up study, Obhi and Hall (2011b) compared human and computer co-actors in a joint task. Participants performed in the classical Libet-clock paradigm, and temporal estimations of actions and outcomes were collected to measure the IBE. Participants were also given feedback on each trial indicating who acted first and caused the tone. The results showed a clear-cut difference between human and computer co-actors. When cooperating with a human co-actor, participants showed a binding effect over their own actions and outcomes but also over those of the co-actor, implying that an implicit SoA has been extended and attributed to the human partner. However, when performing in an identical task with the computer co-actor, no binding was found for the actions and outcomes of the computer. Moreover, in this condition, even self-attribution of agency was overturned. In other words, the SoA was neither extended nor experienced on the individual level.

After partially replicating the results of Obhi and Hall (2011b) in our second study, we were interested to further investigate the factors that might facilitate the attribution of agency to non-human co-actors. Specifically, our aim was to reduce the gap between human and non-human co-actors with regard to action co-representation. In a study from 2012, Berberian and colleagues tested the SoA in an applied setting of an aircraft supervision task. Different levels of system automation were designed to allow participants varying levels of interaction with the computer. By measuring both implicit and explicit agency, results showed a negative correlation between the IBE and subjective agency and the system's level of automation. The more automated the system was (i.e., smaller role for the participant), the weaker the measured binding effect and reported sense of control were.

In another applied experiment by Waytz and colleagues (2014), a driving simulator was used to compare between a normal car, an autonomous car, and an anthropomorphized car with humanlike features. It was found that participants were more inclined to trust a driverless car when augmented with features like name,

gender and voice. When seeming to have a humanlike mind, the autonomous car was perceived as more competent in executing its intended behavior. In study III, we have built on these innovative studies and designed two manipulations for the computer co-actor: one group of participants performed the joint task after training the autonomous car's system, while a second group performed alongside an anthropomorphized version of the car. We were interested in finding which of the two might facilitate agency attribution to the automated system.

2 Research Questions and Hypotheses

This research project investigated theoretically fundamental and applied aspects of the sense of agency. Although gaining growing amount of scientific attention for the past two decades, some central aspects of the SoA remained almost completely unexplored. The three experiments within this project were designed to make the first steps towards filling these gaps and contribute to both human agency research and the design of automated systems.

An abundance of imaging studies (the large majority of which use fMRI) can now be found about the brain structures that underlie self-agency and the ability to discern self and other produced actions and outcomes (for a brief meta-analysis see Sperduti et al., 2011). However, much less is known about the direct link between the neural activity and the central cognitive mechanisms of prediction and retrospective inference assumed to give rise to the SoA. Moreover, not much is known about the influence of the context in which a voluntary action is being performed on the formation of agency. Specifically, it is unclear to what extent the contiguity of preceding action-outcome couplings is decisive for the SoA. Contingency and contiguity are both well-established factors known to shape the context that influences actions in instrumental learning (Shanks & Dickinson, 1991). While contingency has been shown to specifically modulate the perceived time of actions (Moore et al., 2009b), contiguity has not yet been directly tested with regard to the intentional binding effect (but see Moore & Haggard, 2008, partially supporting the role of contiguity in this context). Therefore, in study I the following questions were addressed:

- How are prediction and retrospective inference reflected on the neural level ?
- How do the immediate compared to the long-term context influence the IBE, and which one better accounts for the two mechanisms ?

A study by Jo and colleagues (2014) provided the first direct evidence that the perceived times of actions and outcomes are correlated with the neural activity prior to action execution, that is, the readiness potential (RP). Specifically, the negativation of the early RP was found to be correlated with stronger backward shift of the outcome towards the action. Since the predictive mechanism is assumed to take place prior to

action execution, and taken together with the findings by Jo and colleagues, we expected the RP to be the neural marker best suited to study the contribution of a predictive mechanism. In order to study the inferential mechanisms we had to rely on brain activity that follows the processing of the outcome. As in most other intentional binding studies, we used a tone as the sensory outcome of the action (e.g., Haggard & Clark, 2003). On the neural level, we therefore analyzed the modulation of the auditory evoked potential (AEP), an event-related potential (ERP) that reflects the neural activity associated with the processing of auditory stimuli. AEPs can be analyzed to either express alterations in low-level perception (e.g., the N200 component) as well as to reflect higher cognitive processes (e.g., the P300 component), in which we were interested (Cone-Wesson & Wunderlich, 2003). Accordingly, the following hypotheses were formed:

Hypothesis 1.1: Modulations in amplitudes of the RP and the AEP (specifically the P300) would reflect the contribution of the predictive and retrospective mechanisms to the IBE, respectively.

Hypothesis 1.2: The very recent accumulation of action-outcome couplings, rather than the long-term accumulated amount, would better account for both prediction and retrospective inference.

In study II we aimed to test the attribution of agency to human and non-human co-actors in a more ecologically valid environment. Specifically, we intended to overcome the leap between the highly restricted lab settings of the Libet-clock paradigm (and the IBE) to the real world environment. Additionally, we wanted to find out whether employing task relevant feedback about the source of the action (cf., neutral feedback used in Obhi and Hall, 2011b) would lead to a differentiation between self and other agency attribution when cooperating with the computer co-actor. By generating a more seamless flow between action, outcome, and feedback we assumed participants would regain an implicit sense of control over their self-produced actions when cooperating with the computer. The following questions have been addressed in study II:

- Will human and computer co-actors have the same influence on the SoA when tested in an applied externally valid scenario as when tested in neutral joint task lab setting ?
- Can task-relevant feedback on the action's source create a difference in the implicit SoA of participants when cooperating with a computer co-actor ?

Accordingly, two specific hypotheses were formed:

Hypothesis 2.1: When cooperating with a human co-actor, significant binding effect of actions, outcomes and derived intervals will be found for both self and other produced actions and outcomes (self and extended agency).

Hypothesis 2.2: Due to task-relevant feedback, when cooperating with a non-human co-actor, significant binding effect of actions, outcomes and derived intervals will be found for self-produced but not for the other's actions and outcomes (self but no extended agency).

Study III was designed as a follow-up experiment of study II and its results (discussed in 4.2). As described in the introduction, we were interested in testing two manipulations to a non-human co-actor and aimed to address the following questions:

- What factors might facilitate the attribution of agency to non-human co-actors in a joint task ?
- Can emphasized and specified feedback about the source of the action have an influence on self compared to other produced actions and outcomes ?

Specifically, we focused on the attempt to reduce the gap between the automated system and the human co-actor as it is measured by implicit agency attribution through the IBE. We hypothesized that by supplementing the computer co-actor with humanlike features such as name, gender, voice, face, and body, participants will readily represent the partner's actions and outcomes in the same manner as they do with a human partner. Following a similar line of thought, we have created a simulated training phase in which participants allegedly trained the computer prior to performing

the joint task. After a successful completion of the training phase, the same joint task took place. We argue that by becoming familiar with the inner workings of the computer, the lost sense of control might be regained through generating a humanlike theory of mind. It was left open whether a bottom-up (using external cues) or a top-down (training manipulation) process would be more successful than the other. Additionally, since task relevant feedback was shown to be ineffective with regard to the IBE (in study II), a stronger more specified feedback was designed and we hypothesized that this type of feedback might have a significant influence on implicit pre-reflective agency. Again, two specific hypotheses were formed:

Hypothesis 3.1: Significant intentional binding effect over self-produced, as well as over the computer co-actor's actions and outcomes will be present in cooperating with either a humanized system or with a previously trained system.

Hypothesis 3.2: Specified and emphasized feedback on the trial level about the source of the action will result in significant binding effects for self but not for the other's actions and outcomes (self but no extended agency).

3 General Methodological Approach

3.1 The Libet Clock Paradigm

Measures of the subjective experience of time have existed for many years. One early reported case in experimental psychology is Wilhelm Wundt's complication-clock (Wundt, 1883). Wundt used a pendulum to explore participants' attention to an auditory click. By doing so, objective and subjective temporal measurements of a stimulus were compared (Carlson et al., 2006). About a century later, Benjamin Libet adopted Wundt's chronometric methodology to study volition (1983). Ever since, researchers studying different aspects of motor action have extensively used the Libet-clock paradigm. In study I, we employed Libet's paradigm and recorded brain activity while participants performed the computerized task (see Figure 2.). In the experiment, participants are seated in front of a screen. At the beginning of each trial a small rotating clock appears at the center of the screen. Participants are instructed to look at the center of the clock and not to follow the moving clock hand. Within a given time frame, participants are requested to perform a spontaneous button press ("as the urge occurs") and note the time in which the action took place. At the end of each trial, the participant estimates the time in which she thinks she pressed the button. The estimation method itself varies between experiments: a verbal report, moving the hand clock to the estimated position or typing in numbers from 0-60. While each variation has its own advantages and disadvantages, the susceptibility to biased estimations cannot be completely avoided. Pocket and Miller (2007) have conducted a thorough study about the method of the rotating clock and compared different experimental factors that might influence its reliability and validity. Aspects relating to the physical characteristics of the clock and the instructions given to participants were altered in different conditions. Although some of the aspects were found to lead to variability in responses, the authors concluded that the method could be validated for the use of recording subjective time measurements. One central argument against potential systematic biases like the prior entry effect or dynamic reallocation of attention (Spence and Parise, 2010; Haggard et al., 2002b) is that even if such biases exist, they are to be cancelled out through comparison between baseline and operant conditions.

Figure 2: **The Libet Clock Setup**. The figure shows a sketch of the common Libet-clock setup. Speakers are used to deliver the sound of the tone and brain activity is measured with an EEG cap worn by participants (missing from the sketch).

3.2 The Intentional Binding Effect

Ever since its discovery, the intentional binding effect played a central role in the research of human agency (for a comprehensive review see Moore and Obhi, 2012a). In comparison to subjective explicit measures of agency, the IBE methodology has strong advantages. Subjective reports and introspection, although simple and direct, are highly susceptible to biases and confounds of inter-subjective variability, which cannot be easily controlled for. Moreover, probing subjective reports of agency might influence the experience of pre-reflective agency itself, being the construct of interest (Synofzik et al., 2008a). The IBE paradigm also overcomes some of the major shortcomings arising from the original Libet task. Individual differences in the use of the clock stemming from varying estimation strategies pose no challenge since the IBE is a relative measure (see Figure 3 for a description of the contrasts in the classical paradigm). Clinical research of agency (e.g., with schizophrenia patients) also benefits from the binding effect, as patients who are not always able to deliver introspective verbal reports, can easily perform the computerized task.

In all three experiments comprising this project, we have employed the IBE paradigm, following its original design by Haggard and colleagues (2002a). In study I, the original design was adapted to investigate the predictive and retrospective mechanisms by manipulating the probability of tones in the operant conditions. In studies II and III, the Libet clock and the measurement of the IBE have been adapted to

suit a new applied context. Using a driving simulator to include two actors in a joint task, we have redesigned the classical clock task and the estimation method: gas pedal presses replaced button presses and the clock itself was replaced with a dynamic filling bar. The parameters of the original paradigm (e.g., rotation speed, rotation fashion, visual angle, randomization method etc.) were closely followed to avoid any confounds stemming from a difference in implementation.

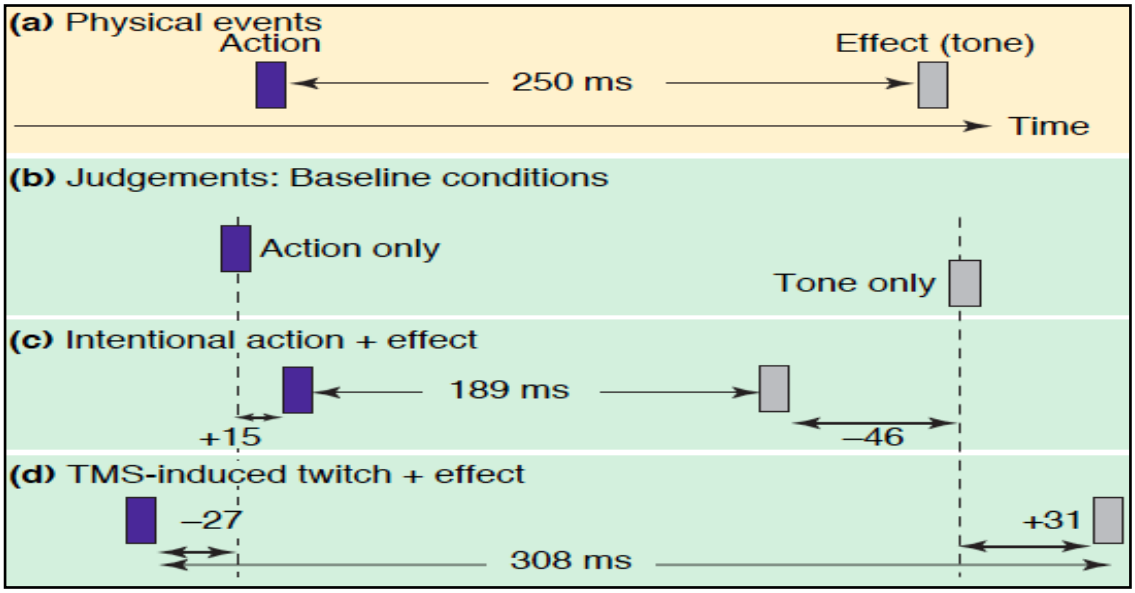

Figure 3: The Intentional Binding Effect **(a)** Participants make self-paced buttons presses, which are followed by a tone after a delay of 250 milliseconds. **(b)** On baseline conditions, participants either press the button or hear a tone, and estimate the time of these events. **(c)** On operant conditions, a button press is followed by a tone and participants estimate the perceived times of these events. Compared to baseline judgements, actions are perceived to occur later and tones are perceived to occur earlier in time. That is, the perceived time interval between action and outcome is reduced. **(d)** When replacing the intentional action with an involuntary movement (induced by TMS), the binding effect is reversed and a "repulsion" effect is found. (Adapted from Haggard, 2005)

3.3 Electroencephalography and Event-Related Potentials

In addition to the behavioral measure, the first of the three studies included in this project also combined brain recordings using electroencephalography (EEG). To further investigate the relation between an action's context and the underlying mechanisms of the SoA, we examined modulations in event-related potentials (ERP) that accompany action and outcome. An overwhelming majority of imaging studies in the field of agency use functional magnetic resonance imaging (fMRI) to

track brain structures and neural networks involved in different aspects of agency attribution (cf., Kang et al., 2015). However, for the purpose of studying the underlying mechanisms of the SoA, we suggest EEG to be a far more suitable option due to its high temporal resolution. Prediction and retrospective inference are two processes assumed to take place before, during and right after the execution of an action and its subsequent sensory outcome. As temporal dynamics of these processes are swift and instantaneous, it seems that the ability to track neural activity on the millisecond level is of crucial importance. Moreover, we were interested in analyzing the neural data in relation to the measurement of the IBE. Although admitting to some variability across different experiments, literature suggests that action and tone binding are found in the range of 20 and 50 milliseconds, respectively (Moore and Obhi, 2012a). As these are considered to be relatively small effect sizes, the importance of an accurate and stable measurement on the neural level is not to be overlooked.

Since button press and tone were used to operationalize action and outcome, we turned to look at the corresponding ERPs: the readiness potential (RP) associated with the button press and the auditory evoked potential (AEP) that accompanies an auditory stimulus. By designing conditions to dissociate the unique contribution of each mechanism (see methods section of study I), as well as the effect of preceding action-outcome coupling, we expected to observe significant modulations in the amplitudes of the ERPs.

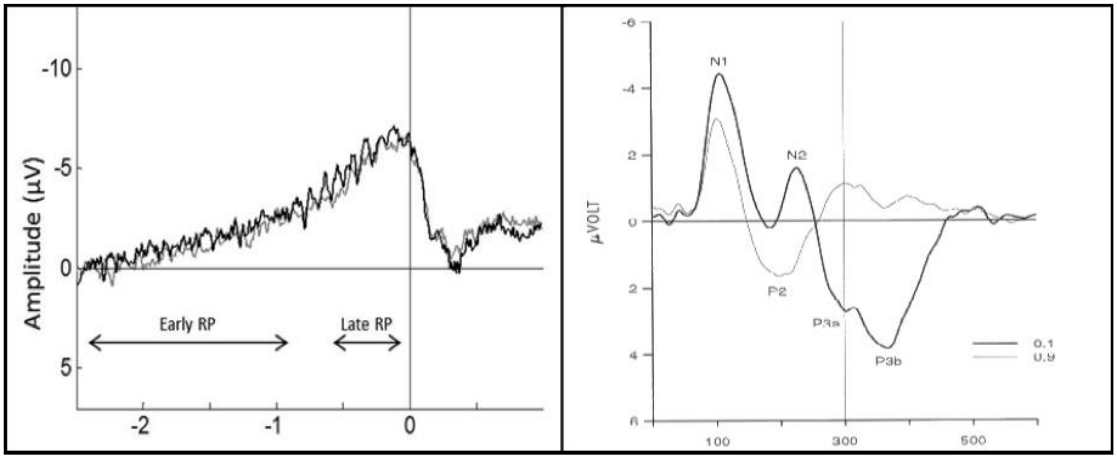

Figure 4: Readiness Potential and Auditory Evoked Potential (Left) The typical readiness potential signal, commonly divided into early and late phases. '0' represents the time of action execution after which the amplitude drops. **(Right)** The darker line represents an auditory evoked potential as can be observed directly after the processing of an auditory stimulus (happening at '0'). Different components have been identified and among them the P300 (divided to P3a and P3b) is of particular interest to our first hypothesis. (Adapted from Jo et al., 2014 and Simons & Perlstein, 1997)

4 Summaries of the Three Experimental Studies

4.1 Study I: The amount of recent action-outcome coupling modulates the mechanisms of the intentional binding effect: A behavioral and ERP study

4.1.1 Background

Recent studies have started to look into the cognitive mechanisms that underlie and modulate the intentional binding effect and the SoA (a.o. see Engbert and Wohlschläger, 2007; Hughes et al., 2012). A multitude of models and theories ultimately converge into two central kinds: predictive and retrospective mechanisms. In the current study we were interested in uncovering the underlying neural correlates of the intricate dynamics of these two mechanisms. For that purpose we have used EEG recordings to track modulations in the corresponding ERPs. Specifically, we expected to observe modulations in the amplitudes of the readiness potential (RP) and the auditory evoked potential (AEP) that will reflect the contribution of the two mechanisms (hypothesis **1.1)**.

Moreover, we were interested in investigating the relationship between an action's context and its temporal perception. Specifically, we wanted to find out how the long-term overall context of the action, as compared to its very recent preceding context, influenced the action binding effect, and which one better accounted for the two mechanisms (i.e., hypothesis **1.2**). To dissociate prediction and inference and compare between immediate and recent contexts, the tone probability was manipulated on operant blocks and a second analysis was devised to inspect the effect of distinct sets of trials.

4.1.2 Methods

The experimental procedure was based on a paradigm by Moore and Haggard (2008) and made use of the Libet-clock and the measurement of the IBE. Each participant was presented with two types of baseline conditions (i.e., action-only and tone-only) and two types of experimental conditions: low and high tone probability. Specifically, the experimental conditions consisted of blocks with either 50% (low) or 75% (high)

probability of trials with a tone following the button press. In each trial, participants (N=24) gave temporal estimations of their actions and outcomes. Data was then analyzed twice: First, long-term accumulation - the judgments of action times were subject to repeated measures ANOVA with probability level (high vs. low) and trial type (tone vs. no-tone) as within-subject variables. Second, recent accumulation – for this analysis we classified single trials according to the accumulated amount of action-tone trials that preceded them (three preceding trials). Each of these preceding trials was then registered, resulting in four different levels of classification. The new division resulted in a new 2x4 factorial design with trial type (tone vs. no-tone) and amount of recent accumulation (None, One, Two, and Three) as within subject factors. To isolate the unique contribution of each mechanism several contrasts were calculated on each analysis (see methods section of study I).

ERPs were calculated in the appropriate regions of interest (ROIs): RP was measured over six lateral electrodes surrounding Cz, where preconscious activation leading to voluntary action is measured (M1, SMA and pre-SMA): FC1, FC2, CP1, CP2, C3 and C4. As participants always pressed the button with their right hand, we were interested in calculating the signal on the contra-lateral side. To do that, activity on three electrodes of each side was averaged. Then, the averaged signal on the right was subtracted from that on the left (Adapted from Eimer, 1998). The AEP was measured around the midline, primary auditory cortex and auditory association areas where the processing of auditory stimuli is most evidently reflected (Picton & Hillyard, 1974): Pz, Cz, Fz, C3, C4, T7, and T8. The amplitude of the RP and AEP on each condition and participant was quantified by calculating the mean signal in the appropriate epochs and baseline corrected.

4.1.3 Results

First analysis (long-term) – The ANOVA revealed no main effect of probability level ($F(1,23) = 0.077$, $p= 0.392$, $\eta2 = 0.003$), but a significant main effect of trial type ($F(1,23) = 5.826$, $p =0.012$, $\eta2= 0.202$) such that regardless of the level of probability, trials with tones showed a significantly stronger action binding than action only trials. There was no significant interaction between probability level and trial type ($F(1,23) =0.167$, $p = 0.343$, $\eta2= 0.007$). The results do not reveal the contribution of a predictive mechanism, as action only trials were not significant on both probability

levels (t(23) = 0.34, p= 0.365; t(23) = 0.36, p= 0.355) and a main effect of probability level was not significant. The electrophysiological analysis therefore focused on the second analysis.

Second analysis (short-term) – The ANOVA revealed a significant main effect of both trial type (F(1,23) = 3.429, p =0.038, Greenhouse-Geisser, η2 =0.130) and amount of recent accumulation (F(3,69) = 3.772, p = 0.007, η2= 0.141) and no significant interaction (F(3,69) = 1.096, p= 0.178, η2= 0.045). The main effect of trial type shows that regardless of the level of recent accumulation, trials with tones showed a significantly stronger action binding than action only trials. The main effect of amount of recent accumulation shows that the action binding gets stronger the more action-tone trials precede a given trial, regardless of its type (with or without tone). Results allowed us to account for both mechanisms. Figure 5 shows the mean shifts from the baseline of the action time judgments according to the second analysis.

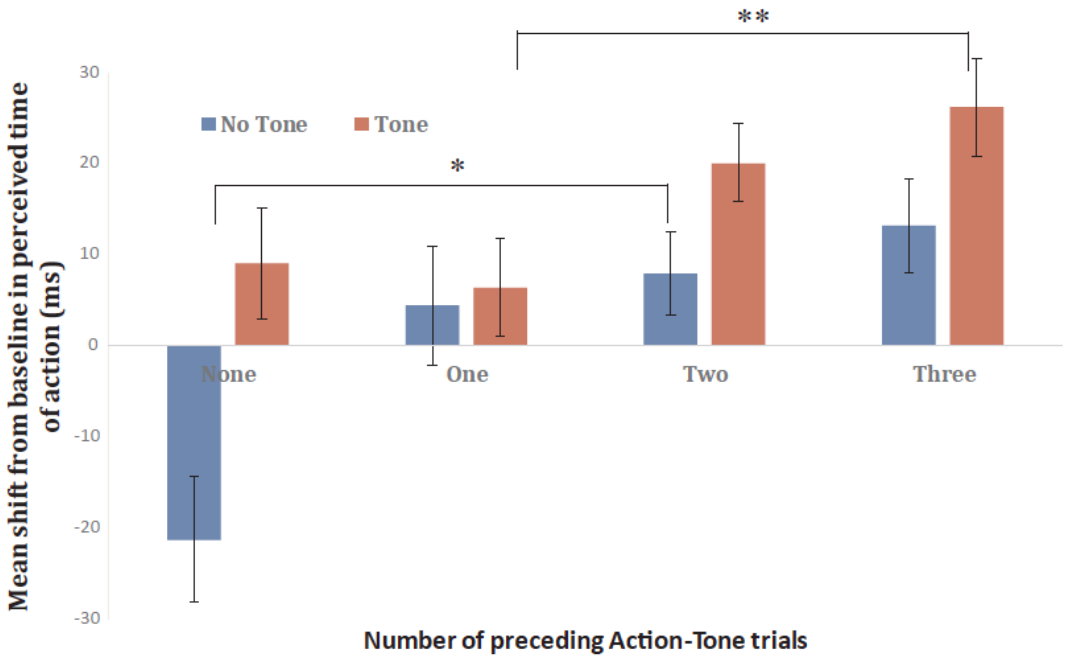

Figure 5: Recent action-outcome coupling. Mean shift from baseline in the perceived time of action for every trial type and amount of recent accumulated action-tone trials. *p < 0.05, **p < 0.01. (Adapted from Goldberg et. al, 2017)

ERP analysis – RP: The a priori ANOVA revealed significant effects in the consecutive time windows between −500 and −275 ms. A second-step one-way ANOVA that was conducted on this bigger time window revealed a significant effect

of amount of recent accumulation (F(1.917, 63.27) =3.076, p =0.034, Greenhouse-Geisser, η2 = 0.219). AEP: The a priori ANOVA revealed a significant main effect of recent accumulation, and a significant interaction between electrode and level of recent accumulation in the consecutive time windows between 300 and 500 ms. A second-step ANOVA that was conducted on this bigger time window revealed a significant main effect of recent accumulation (F(3,69) = 4.610, p= 0.004, η2= 0.295) and a significant interaction between electrode and recent accumulation (F(4.492,103.316) = 2.767, p= 0.016, Greenhouse-Geisser, η2= 0.201). Figure 6 shows the topographic maps to illustrate the difference in activation between the levels of amount of recent experience in the significant time windows of both ERPs.

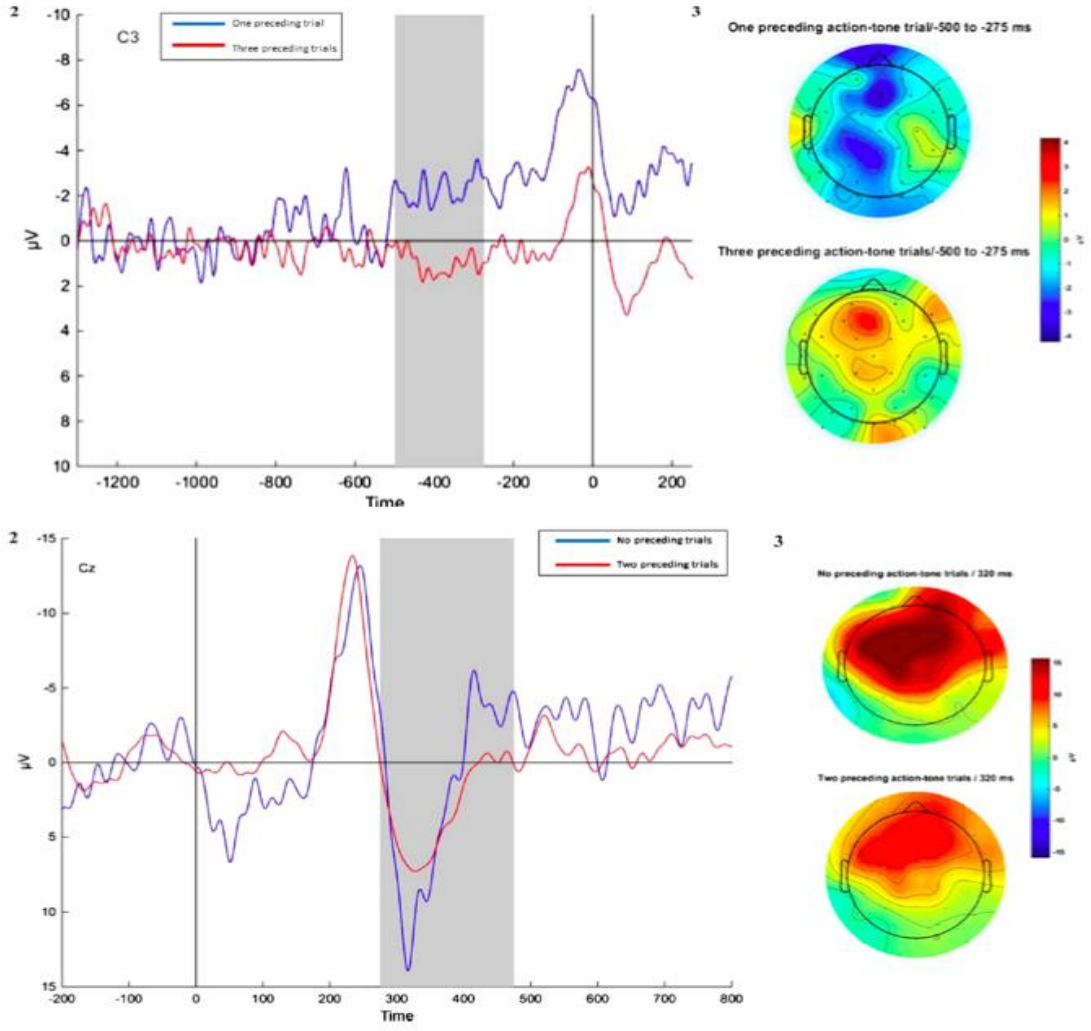

Figure 6: Significant time windows and topographic head maps. Grand averaged RP **(up)** and AEP **(down)**. Significant time windows are marked by the gray rectangle. 0 stands for the time of the button press or tone, respectively. (Adapted from Goldberg et al., 2017)

4.1.4 Discussion

Behavioral results of the first analysis (i.e., long-term accumulation) could not replicate those of Moore and Haggard (2008). Interestingly, our findings mirror the pattern of results of schizophrenic patients in Voss et al., 2010. One crucial point that can potentially underlie the difference between the studies is the confined length of the experiment, and the smaller number of blocks per probability level that was used (see elaboration in final discussion and original article). Moreover, we have shown that the immediate context of our actions has a decisive influence on our temporal perception of the action-outcome relation. Compared to the long-term accumulative context, the recent history of action-outcome coupling could account for both a predictive and an inferential shift of actions towards outcomes, as expected by the intentional binding effect. These results support hypothesis **1.2**. Furthermore, distinct modulations of the RP and the P300 of the AEP (specifically a P3b) were observed. These ERP modulations reflect the implementation of the higher cognitive mechanisms on the neural level and corroborate hypothesis **1.1** (see limitations in final discussion).

4.2 Study II: The Attribution of Agency to Non-Human Co-Actors in a Joint Task: A Driving Scenario

4.2.1 Background

The SoA was recently shown to be modified and extended when cooperating with another partner towards a common goal (Strother et al., 2010). In the current study we wanted to deepen and extend the findings by Obhi and Hall (2011b) who found a difference in agency attribution to human compared to computer co-actor in a joint task. Our central goal was to create an applied, and more ecologically valid task in which the Libet paradigm could be embedded and replicate the pattern of agency attribution in this new environment (hypothesis **2.1**). Furthermore, Obhi and Hall found that feedback about the source of the action had no influence on the implicit measure of agency (only on explicit judgments). Our second goal was to develop a more meaningful, task relevant feedback that would be reflected on the pre-reflective level as measured by the IBE and would lead to self but not extended agency when cooperating with the computer co-actor (hypothesis **2.2**).

4.2.2 Methods

The new joint task was implemented in a driving simulator, whereby participants (N=43) were driving in a two-lane road side-by-side with another driver. The second driver was either a confederate, or a computer program, which was presented to participants as running the system of an autonomous car. The two-lane road was designed to converge at a certain point into a single lane. The goal of the joint task was to avoid an imminent crash at the convergence point by accelerating and overtaking the other driver in advance. Button presses and tones were in the new scenario converted to pedal presses (in order to accelerate) followed by tones signaling a successful avoidance of the crash. Participants performed the task twice with both the confederate and the computer, while unbeknownst to them, the same behavior was employed by the program in both cases (that is, the confederate did not participate). The only difference between the two conditions was the belief manipulation about the partner with which the task was being performed. In the action baseline condition, participants were driving on a straight road towards a highway. Once crossing the highway road sign, the participant was requested to accelerate by pressing the gas pedal once. No tone followed. The participant had to estimate the time of her pedal press. In the tone baseline condition, the participant was driving in a mountainous landscape. A road sign signaled that the driver is about to enter a hazard zone and was requested not to press the gas pedal from that point on in order to decelerate. A tone signaling the end of the danger zone, was presented randomly in an interval of 2 to 6 seconds from crossing the road sign. The participant then estimated its time. The Libet-clock was replaced with a filling bar that was presented during trials on the right side of the windshield. The same bar appeared on the estimations screen where participants had to fill it up to the estimated time point. On each trial, the overtaking of one of the cars served as a feedback about who acted first and elicited the tone and was seamlessly connected to the action-outcome event. The experiment comprised of a 2x2x2x2 multi-factorial, repeated measures design with the following factors: Co-Actor: Human, Computer; Condition Type: Baseline, Operant; Estimated Event: Pedal press, Tone; and Feedback: Self, Other.

4.2.3 Results

To enable a comprehensive and detailed interpretation of the data, a dual analysis was performed: shifts of single events (i.e., action and outcome) and changes in the derived action-outcome intervals across conditions were calculated. The analysis revealed several significant results: a main effect of condition type (operant conditions were always significantly shifted from baseline conditions: $F(1,42)=27.1$, $p<.000$), a main effect of estimated event (stronger binding for tones than for actions: $F(1,42)=8.564$, $p=.006$) and an interaction between the two ($F(1,42)=67.1$, $p<.000$). More interestingly, a significant interaction was found between co-actor, estimated event, and condition type ($F(1,42)=6.141$, $p=.017$). When participants performed the joint task with the confederate, estimates of the onset of the action and tone were both significantly shifted in operant compared to baseline conditions (action forward shift: $t(42)=3.026$, $p=.004$; tone backward shift: $t(42)=-10.3$, $p<.000$). In contrast, when participants performed the joint task with the computer's system, a significant temporal shift was found for the tone, but not for the action (action forward shift: $t(42)=1.568$, $p=.124$; tone backward shift: $t(42)=-7.292$, $p<.000$). See Figure 7 for an illustration of these results. The interval analysis coheres with the results from the single event analysis, insofar as it supports a strong binding effect for the human co-actor (of both action and tone) and a weaker binding effect for the computer co-actor (of only the tone). Finally, feedback showed no influence on the dependent measure in any of the conditions.

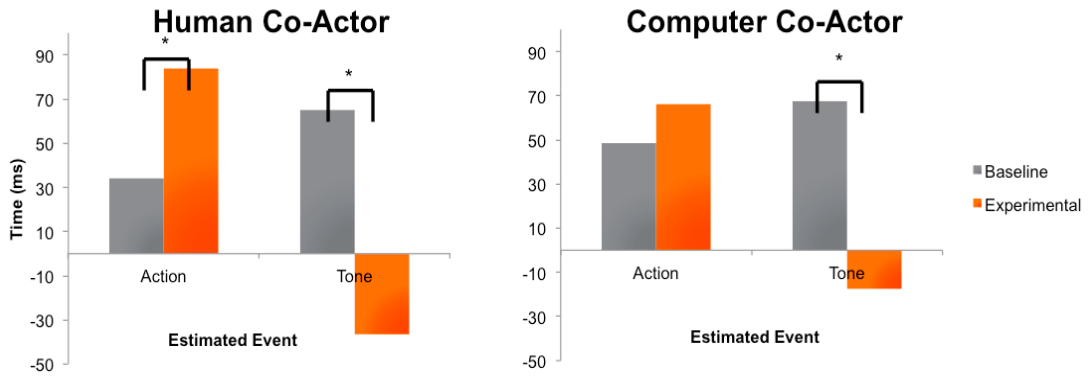

Figure 7: Mean temporal estimations. Temporal estimations of action and tone on both condition types for the two co-actors. (Adapted from Goldberg et al., submitted)

4.2.4 Discussion

Results of the single event analysis replicated those of Obhi and Hall by showing a significant forward action binding when cooperating with the confederate but not with the computer. This is regardless of whose action it was, which was indicated as the effective action. This supports the central thesis that an extended agentic identity is formed when cooperating with another person in a joint task but overturned when cooperating with a computer. The second part of the single event analysis revealed a backward shift of the tone for both co-actors. This result partially failed to replicate the findings by Obhi and Hall, since a backward tone shift was not expected on the human-computer condition. Taken together, the results supported hypothesis **2.1**. One major difference between the current and the original study was the newly designed context in which the joint task was performed. Specifically, in the current study, the joint task departed from the nominal partnerships formed by merely watching the Libet clock. We argue that it was these meaningful contextual differences that account for the extended binding effect not only for the human confederate but also for the automated computer when it comes to the outcomes. While actions are tightly connected to the executing co-actor by being the direct result of a complex internal process of action initiation and preparation, their outcomes are external events only transitively connected to the agent and therefore not susceptible to the type of co-actor with which the task is performed. Finally, task relevant feedback was found ineffective with regard to the IBE, a result that contradicts our prediction (i.e. hypothesis **2.2**) but supports the well-acknowledged distinction between judgment and feeling of agency, parallel to the explicit and implicit levels of the SoA (see Moore et al., 2012b and Dewey and Knoblich, 2014. Find an elaborated discussion in 5.4).

4.3 Study III: Attribution of Agency to Automated Entities: Humanized versus Trained Systems

4.3.1 Background

The current study was a follow-up experiment of study II, and was designed to investigate two types of non-human co-actors: a humanized and a trained computer. In the previous study we showed that human and computer co-actors defer with regard to agency attribution. Specifically, participants were not extending implicit agency over

the actions of the computer co-actor but only when cooperating with the confederate. By manipulating the computer's external features (group one) and by letting participants train its system prior to the joint task (group two), we predicted that a unified agentic identity will emerge in the same manner as with a human co-actor (hypothesis **3.1**). Furthermore, study II provided evidence that task relevant feedback about the source of the action is insufficient in causing any change on the implicit level as measured by the IBE. In the current experiment, we have employed specified and emphasized feedback on the trial level in order to cause a differentiation not only in explicit agency judgments but also in pre-reflective SoA. The effect of feedback was expected to be observed by strong binding over self-produced actions and outcomes with comparison to the ones produced by the co-actor (i.e., hypothesis **3.2**).

4.3.2 Methods

The general experimental design followed that of the previous experiment. Participants (N=40) were assigned to one of two groups and completed the joint task with either the avatar co-actor or with the trained co-actor. For the anthropomorphized computer, a female avatar figure named Iris presented herself to participants as the new autonomous car's system. A set of pre-defined texts was interleaved before, between and after each block. Whenever the avatar talked to participants, her facial and bodily movements were synchronized with her speech, and her figure was displayed at the center of the screen. The avatar commented on the written instructions and reformulated it with her own words or simply commented on the progress of the experiment. By talking to participants during trials, the avatar also served to strengthen the feedback factor. The second group of participants performed the joint task after a simulated training phase: participants were seated in front of the co-actor's set-up, and explained that they are required to train the system before driving alongside it. The system's learning process was described as based on the participant's feedback and through elimination of false behavior. Participants watched the autonomous car driving on the left lane as it performed a set of pre-defined behaviors. The training was divided into 4 stages, each dedicated to train the system a basic driving skill: turning, accelerating, recognizing a street sign and eventually merging into a one-lane road. The required correct behavior was described prior to each step. For each of the four stages we had designed a set of wrong behaviors, deviating from the correct behavior to some degree, from completely wrong to almost correct. These were ordered to

induce a sense of a learning progress. After each trial, participants had to decide whether the observed behavior was correct or incorrect. If incorrect, participants had to further specify which of the four options best describes the false behavior (presented in a multiple choice form). After a successful completion of all four stages, participants moved to their own set-up to do the joint task with the allegedly trained autonomous car.

4.3.3 Results

A dual analysis of single events and derived action-outcome intervals was performed. The single event analysis ANOVA revealed the following significant results: main effect of condition type (operant conditions were significantly shifted from baseline conditions: $F(1,19)=3.56$, $p=.037$), main effect of estimated event (stronger binding for tones than for actions: $F(1,19)=4.53$, $p=.047$), and the interaction between the two ($F(1,19)=44.7$, $p<.000$). Most relevant to our predictions, a two way interaction between co-actor and estimated event was found significant ($F((1,19)=5.55$, $p=0.029$). Post-hoc t-tests showed that when participants performed the joint task with the avatar, temporal estimations of actions were not significantly shifted from the baseline (action forward shift: $t(19)=1.14$, $p=.133$), while those of tones were significantly shifted (tone backward shift: $t(19)=5.14$, $p<.000$). When performing with the trained system, both action (action forward shift: $t(19)=1.92$, $p=.035$) and tone binding (tone backward shift: $t(19)=2.95$, $p=.004$) were significantly shifted. Figure 8 illustrates these findings. Interval analysis showed significant shifts for both co-actors, which only partially reflects the single event analysis. This might be related to the strong binding of tones for both co-actors such that the derived interval turned significant even for the avatar co-actor. Finally, the feedback factor did not reveal the expected effect by showing significant differences between self and other on any of the conditions.

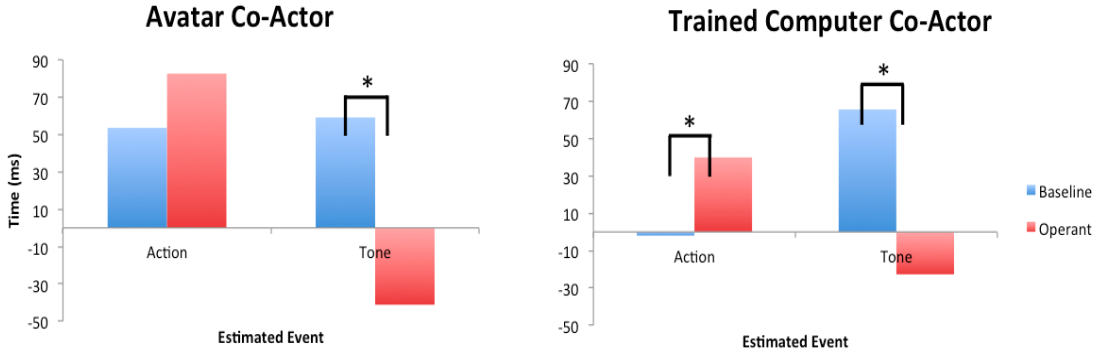

Figure 8: Mean temporal estimations. Temporal estimations of action and tone on both condition types for the humanized and the trained co-actors. (Adapted from Goldberg et al., submitted)

4.3.4 Discussion

In the current study we have shown that a joint task performed after training a computer co-actor results in a similar behavioral pattern to that of cooperating with a human co-actor, supporting hypothesis **3.1**. Participants showed self and extended agency with regard to both actions and outcomes. However, adding anthropomorphic features to the computer did not yield any difference compared to a non-humanized computer and resulted in outcome but not action binding, in contrast to the prediction of hypothesis **3.1**. We argue that although different on the face level, the two manipulations tap into the same underlying process of reducing the gap between human and non-human co-actors through increased sense of similarity to oneself. We appeal to the concept of mind perception in order to offer an explanation to the different results stemming from the two manipulations. In their survey from 2007, Gray and colleagues studied the structure of mind perception and showed that minds are perceived along different dimensions. Their results stress the fact that perceiving an agent as having a mind is not an all-or-none type of process, but rather a complex matter admitting of degrees. Even when restricting the discussion to agency, different aspects can be differentiated (e.g., self-control, ability to plan, decision-making etc.), each of which might be influenced and modulated by different factors. Therefore, it is possible that the training phase manipulation tapped directly into beliefs about action generation, planning and control, while simply presenting human external cues did not connect to the same constructs. Furthermore, our results showed once again that feedback had no influence on the implicit level of agency, and did not support hypothesis **3.2**. Although designed to adapt to specific events and conditions, and while completely accepted by participants as shown by a debriefing session, no difference was found between temporal estimations of self and other produced actions and outcomes. An elaborated discussion of the distinction between judgment and feeling of agency follows in the next section and may shed more light on our results.

5 General Discussion

5.1 Summary of Results

The objective of the current research project was twofold. First, we conducted a study in order to investigate the neural correlates and factors that modulate the underlying mechanisms of the SoA. Second, we were interested in testing the attribution of agency in an applied framework of a joint task. Specifically, we compared different types of co-actors and focused on the difference between human and non-human partners. A summary of the results, tested against hypotheses, will now be presented. Selected findings will then be discussed and integrated, going beyond the discussion of each individual study in the previous section. Finally, I would like to point to some theoretical and methodological limitations and provide suggestions for future research.

Study I corroborated our two central hypotheses (**1.1** and **1.2**). First, no-tone trials with more recent accumulated action-tone couplings were reflected in significantly reduced negative amplitudes of the RP compared to those with less accumulated couplings. In the absence of a tone, the difference in action binding effect and RP amplitudes is attributed to the different levels of recent accumulated action-tone couplings that preceded them. This is considered as supporting the expression of a predictive mechanism. Moreover, our findings point to the relation between a stronger temporal binding of actions and a diminished P300 component, which is modulated through higher levels of preceding action-outcome couplings. Tone trials displayed a significantly smaller positive amplitude around the P300 component of the AEP for trials with more compared to less (or none) recent accumulated couplings. The contribution of an inferential component to the action binding effect is therefore reflected in the overall stronger binding in tone trials compared to no-tone trials across all levels of recent action-outcome couplings. On the neural level, however, a direct contrast between tone and no-tone trials was not possible due to the fact that a comparison between AEPs had to include only trials with tones. Therefore, a straightforward conclusion cannot be drawn in this regard. Second, the analysis based on the long-term accumulation of action-outcome coupling could not support a predictive account and failed to replicate the findings by Moore and Haggard (2008).

Significant temporal shifts were found only for trials with tones compared to trials without tones, but not for higher compared to lower tone probability levels. In contrast, a consistent positive relation between the amount of recent accumulated action-tone couplings and the perceived shift in action time was found, such that the more action-outcome trials preceding a single trial, the stronger the temporal action binding. This was found to be true for no-tone and tone trials, a fact that suggests a common ground (i.e., the influence of an immediate context) for the predictive and inferential components.

The results of study II supported hypothesis **2.1**, but disproved hypothesis **2.2**. When driving alongside a confederate, participants showed significant binding for both actions and outcomes, regardless of who performed the effective action. That is, self as well as extended agency were found when cooperating with the human co-actor. These results are in line with Obhi and Hall's (2011b) findings that employed the joint task using the original Libet paradigm. However, in contrast to our prediction, a cooperation with the computer did not result in a complete inversion. Surprisingly, participants showed binding for self and other produced outcomes, a finding that is possibly related to the newly applied environment. Action binding followed the expected pattern, and was not significant when cooperating with the computer. Moreover, feedback had no influence on the binding of actions and effects, expressed by no differentiation in binding of self and other actions and outcomes. Our results suggest that while actions might be more susceptible to the type of co-actor, outcomes seem to be more affected by the agentic unity created through a meaningful joint task.

Finally, in study III we compared two types of non-human co-actors in a joint task identical to the one presented in study II. Results revealed that a joint task performed with a computer after a simulated training phase, has caused participants to attribute it with agency for both actions and outcomes. At the same time, when a computer was augmented with humanlike features (e.g., name, gender, voice etc.), no such effect was found and the results mirrored those of a normal computer in the previous study. These findings support hypothesis **3.1.** Additionally, specified and strengthened feedback implemented by embedding the avatar at the end of each action-outcome sequence showed no effect on the implicit measurement of agency. This result is inconsistent with hypothesis **3.2** but is in line with previous findings that point to the futility of explicit feedback in modulating the pre-reflective sense of agency and its attribution to others.

5.2 Optimal Cue Integration

In our first study we provided evidence for the contribution of predictive and retrospective mechanisms to the formation of the SoA. Behavioral results demonstrated that the immediate recent context of a given action, in terms of the amount of preceding action-outcome couplings, have a decisive influence on its temporal perception. Both prediction (no-tone trials) and retrospective inference (tone trials) were accounted for. Electrophysiological data lent further support to this view by providing evidence about the modulation of specific ERPs.

While the research of human agency is now playing a more prominent role in the cognitive sciences, the experimental field is still in its infancy and theoretical progress has not been exhausted. One such important step has been recently made with regard to the underlying mechanisms of the SoA. Although previously thought that the experience of agency could be explained by relying on either one of the discussed accounts, it is becoming clear that the two are not mutually exclusive. In a study from 2009(a), Moore and colleagues formalized a new framework to account for the emergence of the SoA. To this day, this is the only theoretical endeavor made to bring together and explain findings of different studies. This new theoretical framework is called 'Optimal Cue Integration' and relies on a statistical solution similar to the maximum likelihood estimation rule (Ernst & Banks, 2002). Generally, according to the theory's central premise, "predictive and postdictive processes each serve as authorship cues that are continuously integrated and weighted depending on their availability and reliability in a given situation" (Synofzik et al., 2013, p. 1). In other words, an integration of sensorimotor and cognitive signals, such as internal motor signals, priming effects and even an outcome's valence are all taken together and processed according to their relative prominence in a specific situation. In terms of the intentional binding effect, action and outcome are themselves two central cues that provide information related to the timing of the other event (Wolpe et al., 2013). The resulting temporal perception as measured by the IBE is therefore a weighted average determined by the reliability of action and outcome. The reason for not relying solely on intrinsic efferent signals (being immediate accessible cues) is that "no single information signal seems to be powerful enough to convey an adequate representation of agency under all everyday conditions" (Synofzik et al., 2009, p. 1067).

It is interesting to consider our first study in light of this integrative framework. One important point in this regard is that our design did not enable us to measure the interaction between the two mechanisms. This shortcoming could be overcome if the reliability of action and tone was intentionally manipulated in advance and predictions can be made about the effects each one of them will have on the binding effect. Moreover, the current study focused on the binding of actions, and did not look into the binding of tones. A different paradigm is required to manipulate the action's reliability and measure its effect on outcome's timing. Having said that, the findings of the first study strengthen the established acknowledgment that both prediction and retrospective inference take part in the formation of an implicit SoA. An interesting step for future research of cue integration would be to find out how the suggested neural correlates (i.e., late RP and the P300) interact as the reliability of different cues is being manipulated.

5.3 Action Co-Representation

Since cognitive neuroscientists have predominantly focused on the study of minds and brains in individuals, rather than in a social context, the exploration of the processes that govern joint action have been left somewhat forsaken. Yet, over the last two decades, more studies have turned the spot in this direction and investigated action and perception in a social context (for a review see Knoblich and Sebanz, 2006 and Sebanz et al., 2006). Our second study showed that participants are inclined to attribute agency to their human partners in a joint task, while this is not the case for a non-human co-actor. Moreover, in a third follow-up study we found that participants who had the chance to train the computer's system prior to cooperating with it, exhibited the same pattern of agency attribution to the co-actor. We suggest that these results can be accommodated within the larger framework of action co-representation. According to Sebanz and colleagues (2006), a successful joint action depends on the ability to share representations, predict actions as well as integrate the predicted outcomes of one's own and the other's actions. Such co-representation might rely on the close link between action and perception as suggested by the common-coding hypothesis: perceiving and performing an action will result in the activation of the same representations through the motor areas. Evidence to the common-coding hypothesis comes from the discovery of the mirror-neurons system, showing that "when

individuals observe an action done by another individual their motor cortex becomes active, in the absence of any overt motor activity" (Rizzolatti and Craighero, 2004, p. 174). In studies II and III, however, participants had no direct visual or auditory contact with the co-actor (the set-ups were separated by a curtain and participants were given earplugs). Therefore, an explanation of the results by appealing to co-representation of the co-actor's actions cannot be based on direct observation. Interestingly, several studies have shown that by simply sharing a task, other mechanisms than direct action observation can result in action co-representation. Specifically, it is the knowledge about the other's task and the conditions under which it will perform a given action, that are important for forming a shared representation. One example comes from a study by Sebanz and colleagues (2005), in which pairs of participants performed together in a reaction time task, each responding to a different dimension of the same stimulus. The results showed that reaction times were slowed on trials in which both co-actors had to respond compared to trials in which only one of them was required to react. Moreover, the same effect was found even when participants could not directly observe the co-actor's actions but were merely aware of their task. Slowed reaction times indicate that the stimulus created some sort of cognitive conflict as two task rules were activated simultaneously. A simultaneous activation of a second task rule to which one is not required to attend is a good example of representing the partner's action (the social Simon effect is a closely related paradigm. For a review see Dolk et al., 2014).

This line of research and the broad theoretical framework of action co-representation provide a satisfying context to consider our results. However, it remains a challenge for future research to determine in what way these neurocognitive processes relate to the IBE and cause a modulation in temporal perceptions of actions.

5.4 Implicit and Explicit SoA

A common objective of both study II and III was to test whether feedback given to participants about the source of the action (i.e., who acted first and caused the outcome) would have an influence on the implicit measure of agency. We found that neither task relevant nor specified and strengthened types of feedback had any influence on pre-reflective agency, as indicated by the IBE.

Our results, although disproving our original predictions, are in line with the findings by Obhi and Hall (2011b) who used patches of color to indicate the source of an action (that is, neutral feedback), and found no influence on the binding effect. Taken together, this is a compelling amount of evidence that supports the dissociability of the implicit and explicit aspects of agency. In a paper from 2008(b), Synofzik and colleagues provide a comprehensive and systematic account of the central features of the acting self. Among other constructs, the authors distinguish between different fundamental dimensions of the SoA: the level of feeling, judging, and ascribing responsibility. For the sake of brevity, we shall not discuss the third of the three mentioned levels, as it has no direct bearings on the current topic. The first and most elementary of the three dimensions is the feeling of agency (FoA). A FoA is characterized as a non-conceptual, stable, and abstract feeling of being the agent that caused a given action. It is an implicit feeling, which can be accessed without mediation of any linguistic representation. The second high-level dimension is named a judgment of agency (JoA). A JoA is a propositional interpretative representation of the self as being the agent causing a specific action. It is, therefore, an explicit conceptualization of the self in contrast to other agents.

The relation between the feeling and the judgment of agency is not straightforward. On the one hand, the FoA is normally a necessary condition for the JoA. For example: my belief that I switched on the lights depends on my experience of reaching for the light switch (Haggard and Tsakiris, 2009). On the other hand, the FoA is not a sufficient condition for the JoA: an explicit JoA requires additionally that the action's outcome will be monitored by the agent. Using the same example, it is only when I see that the lights have come on would I judge that *I* have switched them on. It has been empirically shown that the two dimensions, despite being closely related, are separable and may even tap into different underlying processes (e.g., Moore et al., 2012b; Kumar and Srinivasan, 2013; Dewey and Knoblich, 2014).

The social context, where multiple co-actors cooperate towards a common goal, is one of these situations in which the two dimensions can be separated and even contradict each other. The results of our studies demonstrate that while participants trust the feedback and form explicit beliefs on its basis, the implicit level remains independent and reflects extended agency where it is not to be found and vice versa. While we had the expectation that different more prominent types of feedback would make a difference, we were proved wrong in this regard.

5.5 Limitations of the Project and Suggestions for Further Research

As is the case in most experimental studies, the current research project tackled a variety of hurdles. While some could be resolved and managed, others were not, and as some were foreseen, others were only revealed during the progress of the project or in hindsight. I shall mention some of these central limitations and suggest steps to be taken in future research (other limitations are found in the original research articles).

Two major drawbacks should be noted with regard to study I: First, as EEG recordings accompanied the behavioral task, we intentionally shortened the length of the session by applying a lower number of blocks per participant. Therefore, the possibility that predictions were not formed over the course of the session due to its length cannot be completely disregarded. Second, the behavioral paradigm designed to measure the unique contribution of a retrospective mechanism relies on a contrast between tone and no-tone trials. This contrast became problematic when measuring modulations in the AEP, since by definition, the ERP is based on the perception of an auditory stimulus. As a result, our conclusion about the retrospective mechanism connecting the behavioral and neural levels remains speculative and requires further support. One possibility to deal with the limitation posed by including EEG is to drop blocks in which tone times are being estimated (assuming, of course they are not the focus of interest). In that case, special care needs to be taken to make sure that participants do not ignore the tones altogether (as they are still presented in operant blocks, but not judged). Making use of unique catch trials inserted between normal trials might be beneficial. Concerning the second limitation, future research interested in measuring the AEP as a neural correlate of an inferential mechanism has to replace the all-or-none comparison made in Moore and Haggard's design (2008), in favor of a gradual presentation of tones (as, for example, in Wolpe et al., 2013).

A great effort was invested in the implementation of studies II and III in order to ensure that the parameters of the original Libet-clock and IBE paradigm were closely followed. However, since one of our central goals was to implement Obhi and Hall's (2011b) joint task in an ecologically valid environment, the new setting forced us to omit explicit agency judgments. Instead, we used a debriefing session to question participants about the way they perceived the feedback given during trials. This is of course not ideal and cannot replace formal agency ratings. Future studies would benefit from a design that can integrate explicit ratings in an uninterrupted fashion.

5.6 Conclusions

The intellectual endeavor to explain the self remains one of the greatest challenges of human reason. The sense of agency plays a fundamental role in this ongoing challenge and poses a variety of intriguing questions and problems for both scientists and philosophers. It seems that many more years of combined effort are required for a comprehensive framework to emerge and integrate all aspects of this elusive concept.

In the current research project we have looked into the underlying neurocognitive mechanisms of the SoA and extended existing findings by showing that both the RP and the AEP reflect the implicit SoA, as measured by the IBE. Unlike previous studies, we did not replicate findings about the formation of a predictive account through long-term accumulation of action-outcome couplings. In contrast, we showed that it is the immediate context of the action that can account for both mechanisms.

Furthermore, we have investigated the new exciting interface between the field of human-computer-interaction (HCI) and the SoA. Our second and third studies used a driving scenario to couple participants with human and non-human partners in a joint task. We found that whilst outcomes are not influenced by the type of co-actor and extended in both cases, actions are only extended when cooperating with a human co-actor. In a follow-up study, a simulated training phase for the computer co-actor prior to the joint task showed to close the gap between human and computer co-actors. The formerly trained computer was attributed agency in the same manner as the human partner. However, simply adding humanlike features to the computer did not result in the same effect. Finally, feedback about the source of the action influenced only explicit beliefs but not the implicit measurement. The findings of the current project push the empirical field of human agency one step closer to a fuller and deeper understanding of its complex subject matters.

In the years to follow, theoretical and empirical progress on the sense of agency may come from new exciting directions. One such development can follow the use of virtual and augmented reality as a tool to take a closer look at first person experiences. Once this instrument will be well established within the field, its combination with imaging methods will open up a myriad of possibilities and new research questions. Complex interactions like the one investigated in this project may be studied in much higher resolution and limitations of the type tackled here can be surmounted.

References

Aarts, H., Custers, R., and Wegner, D.M. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition*. 14(3), 439-458. doi: 10.1016/j.concog.2004.11.001

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617-645. doi: 10.1146/annurev.psych.59.103006.093639

Berberian, B., Sarrazin, J-C., Le Blaye, P., and Haggard, P. (2012). Automation Technology and Sense of Control: A Window on Human Agency. *PLoS ONE*, 7(3): e34075. doi:10.1371/journal.pone.0034075

Blakemore, S. J., and Frith, C. (2003). Self-awareness and action. Current Opinions in Neurobiology, 13(2), 219–224. doi: 10.1016/S0959-4388(03)00043-6

Carlson, T.A., Hogendoorn, H., and Verstraten, F.A.J. (2006). The speed of visual attention: What time is it? *Journal of Vision*. 6, 1406-1411. doi: 10.1167/6.12.6

Cone-Wesson, B., Wunderlich, B. (2003). Auditory evoked potentials from the cortex: audiology applications. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 11, 372-377. doi: 10.1097/00020840-200310000-00011

Coyle, D., Moore, J., Kristensson., P.O., Fletcher, P.C., and Blackwell, A.F. (2012). I did that! Measuring Users' Experience of Agency in their own Actions. In *CHI, ACM Conference on Human Factors in Computing Systems*, (Austin, Texas, USA), 2025–2034

David, N., Newen, A., and Vogeley, K. (2008). The ''sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*. 17(2), 523-524. doi: 10.1016/j.concog.2008.03.004

de Haan, S., and de Bruin, L. (2010). Reconstructing the minimal self, or how to make sense of agency and ownership. *Phenomennology and the Cognitive Sciences*. 9(3), 373-396. doi: 10.1007/s11097-009-9148-0

Dewey, J.A., and Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing? *PLoS ONE*. 9(10). doi: 10.1371/ journal.pone.0110118

Dolk, T., Hommel, B., Colzato, L.S., Schütz-Bosbach, S., Prinz, W., and Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. *Frontiers in Psychology*, 5(974), doi: 10.3389/fpsyg.2014.00974

Eimer, M. (1998). The lateralized readiness potential as an on-line measure of central response activation processes. *Behavior Research Methods, Instruments, & Computers*. 30(1), 146-156. doi: 10.3758/BF03209424

Engbert, K., and Wohlschläger, A. (2007). Intentions and expectations in temporal binding. *Consciousness and Cognition*. 16, 255–264. doi: 10.1016/j. concog.2006.09.010

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*. *415*(6870), 429–433. doi: 10.1038/415429a

Farrer, C., Bouchereau, M., Jeannerod, M., and Franck, N. (2008). Effect of distorted visual feedback on the sense of agency. *Behavioural Neurology.* 19(1-2), 53–57. doi: 10.1155/2008/425267

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Science*, 4(1), 14-21. doi: 10.1016/S1364-6613(99)01417-5

Gray, H.M., Gray, K., and Wegner, D.M. (2007). Dimensions of Mind Perception. *Science*. 315, 619. doi: 10.1126/science.1134475

Haggard, P., Aschersleben, G., Gehrke, J., & Prinz, W. (2002a). Action, binding, and awareness. In W. Prinz & B Hommel (Eds.) *Common Mechanisms in Perception and Action* (Attention and Performance, Vol. XIX, 266-285). Oxford, UK: Oxford University Press.

Haggard, P., Clark, S., & Kalogeras, J. (2002b). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382–385. doi:10.1038/nn827

Haggard, P., Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition*, 12, 695-707. doi: 10.1016/S1053-8100(03)00052-7

Haggard, P. (2005). Conscious intention and motor cognition. *TRENDS in Cognitive Sciences*. 9(6), 290-295. doi: 10.1016/j.tics.2005.04.012

Haggard, P., and Tsakiris, M. (2009). The experience of agency: Feelings, judgments and responsibility. *Current Directions in Psychological Science*. 18(4), 242-246. doi: 10.1111/j.1467-8721.2009.01644.x

Haggard, P. (2017). Sense of Agency in the Human Brain. *Nature Reviews. Neuroscience*. 18(4),196-207. doi: 10.1038/nrn.2017.14

Hughes, G., Desantis, A., and Waszak, F. (2012). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*. 139(1), 1133–1151. doi: 10.1037/a0028566

James, W. (1890). *The Principles of Psychology*. New York: Dover Publications (reprinted 1950).

Jo, H.G., Wittmann, M., Hinterberger, T., & Schmidt, S. (2014). The readiness potential reflects intentional binding. *Frontiers in Human Neuroscience*, 8:421. doi: 10.3389/fnhum.2014.00421

Kang, S.Y., Im, C.-H., Shim, M., Nahab, F.B., Park, J., Kim, D.-W., Kakareka, J., and Mileta, N. (2015). Brain Networks Responsible for Sense of Agency: An EEG Study. *PLoS ONE*. 10(9), e0137769. doi: 10.1371/journal.pone.0137769

Kihlstrom, J. F., Beer, J. S., and Klein, S. B. (2003). Self and identity as memory. In M.R.Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 68-90). New York: Guilford Press.

Klein, S. (2002). Libet's Research on the Timing of Conscious Intention to Act: A Commentary. *Consciousness and Cognition*. 11(2), 273-279. doi: 10.1006/ccog.2002.0557

Knoblich, G., and Sebanz, N. (2006). The social nature of perception and action. *Current Directions in Psychological Science*. 15(3), 99–104. doi: 10.1111/j.0963-7214.2006.00415.x

Kumar, D., and Srinivasan, N. (2013). Hierarchical control and sense of agency: Differential effects of control on implicit and explicit measures of agency. In Proceedings of 35th Annual Meeting of the Cognitive Science Society, Berlin, Germany

Libet, B., Gleason, C.A., Wright, E.W., and Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*. 106(3), 632-642. doi: 10.1093/brain/106.3.623

Limerick, H., Coyle1, D., and Moore, J.W. (2014). The experience of agency in human-computer interactions: a review. *Frontiers in Human Neuroscience*, 8(643), doi: 10.3389/fnhum.2014.00643

Locke, J., (1694). *An Essay Concerning Human Understanding*, ed. P. Nidditch, Oxford: Clarendon Press (original work, 2nd ed., partly reprinted in Perry, 1975).

Moore, J.W., and Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, 17, 136-144. doi: 10.1016/j.concog.2006.12.004

Moore, J. W., Wegner, D. M., and Haggard, P. (2009a). Modulating the sense of agency with external cues. *Consciousness and Cognition*, 18(4), 1056-1064. doi:10.1016/j.concog.2009.05.004

Moore, J.M., Lagnado, D., Deal, D.C., and Haggard, P. (2009b). Feelings of control: Contingency determines experience of action. *Cognition*, 110, 279-283. doi: 10.1016/j.cognition.2008.11.006

Moore, J.W., and Obhi, S.S. (2012a). Intentional binding and the sense of agency: A review. *Consciousness and Cognition*. 21(1), 546-561. doi: 10.1016/j.concog.2011.12.002

Moore, J.W., Middleton, D., Haggard, P., and Fletcher, P.C. (2012b). Exploring implicit and explicit aspects of sense of agency. *Consciousness and Cognition*. 21(4), 1748-1753. doi: 10.1016/j.concog.2012.10.005

Nichols, S. and M. Bruno (2010). Intuitions about Personal Identity: An Empirical Study, *Philosophical Psychology*. 23(3), 293–312. doi: 10.1080/09515089.2010.490939

Niedenthal, P., Barsalou, L. W., Winkielman, P., Kraut-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review, 9,* 184-211. doi: 10.1207/s15327957pspr0903_1

Obhi, S.S., and Hall, P. (2011a). Sense of agency and intentional binding in joint action. *Experimental Brain Research*, 211, 655–662, doi: 10.1007/s00221-011-2675-2

Obhi, S.S., and Hall, P. (2011b). Sense of agency in joint action: influence of human and computer co-actors. *Experimental Brain Research*, 211, 663–670, doi: 10.1007/s00221-011-2662-7

Pockett, S., Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition*. 16 (2), 241-254. doi: 10.1016/j.concog.2006.09.002

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annual Review Neuroscience*. 27, 169-192. doi: 10.1146/annurev.neuro.27.070203.144230

Searle, J. (2001). Free Will as a Problem in Neurobiology. *Philosophy*, 76(4), 491-514. doi: 10.1017/S0031819101000535

Sebanz, N., Knoblick, G., and Prinz, W. (2005). How to share a task: corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance*.31(6), 1234-1246. doi: 10.1037/0096-1523.31.6.1234

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *TRENDS in Cognitive Sciences*. 10(2), 70–76. doi: 10.1016/j.tics.2005.12.009

Sedikides, C. and Skowronski, J. J. (1997). The symbolic self in evolutionary context. *Personality and Social Psychology Review, 1,* 80-102. doi: 10.1207/s15327957pspr0101_6

Shanks, D.R., and Dickinson, A, (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, 19(4), 353-360. doi: 10.3758/BF03197139

Shoemaker, S. (1984). Personal Identity: A Materialist's Account, in Shoemaker and Swinburne, *Personal Identity*, Oxford: Blackwell.

Simons, R.F., and Perlstein, W.M. (1997). A tale of two reflexes: an ERP analysis of prepulse inhibition and orienting. In: Simons, R.F., and Lang, P.J., editors. *Attention and Orienting: Sensory and Motivational Processes*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. pp. 229–255.

Spence, C., and Parise, C. (2010). Prior-entry: A review. *Consciousness and Cognition*. 19(1), 364-79. doi: 10.1016/j.concog.2009.12.001

Sperduti, M., Delaveau, P., Fossati, P., and Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: a brief review and meta-analysis. *Brain Structure and Function*. 216(2), 151-157. doi: 10.1007/s00429-010-0298-1

Strother, L., House, K.A., Obhi, S.S. (2010). Subjective agency and awareness of shared actions. *Conscious and Cognition*, 19(1), 12–20. doi: 10.1016/j.concog.2009.12.007

Synofzik, M., Vosgerau, G., and Newen, A. (2008a). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and* Cognition. 17(1), 219–239. doi: 10.1016/j.concog.2007.03.010

Synofzik, M., Vosgerau, G., and Newen, A. (2008b). I move, therefore I am: A new theoretical framework to investigate agency and ownership. *Consciousness and Cognition*. 17(2), 411-424. doi: 10.1016/j.concog.2008.03.008

Synofzik, M., Vosgerau, G., and Lindner, A. (2009). Me or not me – An optimal integration of agency cues? *Consciousness and Cognition*. 18(4), 1065-1068. doi: 10.1016/j.concog.2009.07.007

Synofzik, M., Vosgerau, G., and Voss, M. (2013). The experience of agency: an interplay between prediction and postdiction. *Frontiers in Psychology*. 4(127). doi: 10.3389/fpsyg.2013.00127

Tsai, C-C., Kuo, W-J., Jing, J-T., Hung, D.L., and Tzeng, O.J.-L (2006). A common coding framework in self–other interaction: evidence from joint action task. *Experimental Brain Research*, 175, 353–362, doi: 10.1007/s00221-006-0557-9

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., and Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain*. 133, 3104-3112. doi: 10.1093/brain/awq152

Waytz, A., Heafner, J.and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113-117. doi: 10.1016/j.jesp.2014.01.005

Wegner, D.M., and Whetley, T. (1999). Apparent Mental Causation: Sources of the Experience of Will. *Ameerican Psychologist*, 54(7), 480-492. doi: 10.1037/0003-066X.54.7.480.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Basil Blackwell Publishing.

Wohlschläger, A., Haggard, P., Gesierich, B., and Prinz, W. (2003). The Perceived Onset Time of Self- and Other-Generated Actions. *Psychological Science*, 14(6), 586-591, doi: 10.1046/j.0956-7976.2003.psci_1469.x

Wolpe, N., Haggard, P., Siebner, H.R., and Rowe, J.B. (2013). Cue integration and the perception of action in intentional binding. *Experimental Brain Research*. 229, 467-474. doi: 10.1007/s00221-013-3419-2

Wundt, W. M. (1883). *Philosophische Studien*. Leipzig: Wilhelm Engelmann.

Original Research Articles

Consciousness and Cognition xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

The amount of recent action-outcome coupling modulates the mechanisms of the intentional binding effect: A behavioral and ERP study

Michael Goldberg[a,b,*], Niko Busch[c,d], Elke van der Meer[a,b]

[a] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10117 Berlin, Germany
[b] Institute of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany
[c] Institute of Psychology, University of Münster, Fliednerstraße 21, 48149 Münster, Germany
[d] Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Münster, Germany

ARTICLE INFO

Keywords:
Sense of agency
Intentional binding effect
Prediction and inference
Readiness potential
Auditory evoked potential
Action-outcome coupling

ABSTRACT

Our everyday interactions depend on the ability to maintain a feeling of control over our bodily actions, that is, the sense of agency. The intentional binding effect – a perceived temporal shortening between voluntary actions and sensory outcomes – has been shown to implicitly measure agency. We investigated the effect's underlying mechanisms: prediction and retrospective inference. First, long-term and recent action-outcome coupling were compared. Second, brain activity was recorded to uncover the neural correlates of the two mechanisms. Our results show that the recent accumulation of action-outcome coupling, but not that of a long-term accumulation, is correlated with the binding effect of actions and accounts for both mechanisms. Temporal action binding was reflected in both the readiness potential and the auditory evoked potential. The results shed new light on our understanding of the influence that immediate context of an action has on its temporal binding and the neural substrate of human agency.

1. Introduction

In recent years, a growing scientific interest in the Sense of Agency (SoA) has yielded an ample amount of data that sheds new light on this essential component of our everyday actions and interactions. The SoA refers to the experience of initiating and controlling actions and their sensory effects. The seamless ongoing feeling of motor control entails, in turn, a stronger sense of causality over the external sensory outcomes in the environment. For a long period of time, the SoA was mainly the focus of philosophers' discussions about free will, self-identity, or consciousness. It has only been during the last decade that newly improved methods and better theoretical frameworks have permitted thorough scientific investigation of the topic within the field of cognitive neuroscience (for a review see, Moore & Obhi, 2012 and David, Obhi, & Moore, 2015).

A key player in this field is the "intentional binding effect", which has been proposed to serve as an implicit and objective measure of the SoA. The effect, originally described by Haggard, Clark, and Kalogeras (2002), is a perceived temporal contraction of the interval between voluntary actions and their sensory outcomes. More specifically, when voluntary actions are closely coupled with sensory outcomes, the actions are perceived later, and the outcomes earlier, as compared to when both appear on their own. This temporal binding occurs for self-initiated but not passive or triggered actions, and is therefore a marker of the SoA. In contrast with verbal, explicit reports of agency, the effect is generally considered to measure a pre-reflective, non-verbal SoA. However, it is to be

* Corresponding author at: Institute of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany.
E-mail address: michael.goldberg@hu-berlin.de (M. Goldberg).

http://dx.doi.org/10.1016/j.concog.2017.07.001
Received 10 August 2016; Received in revised form 26 June 2017; Accepted 3 July 2017

noted that an ongoing debate still exists as to the exact nature of the measured construct and the relation between different explicit and implicit measures as well as between different implicit measures, like for example sensory attenuation (cf., Dewey & Knoblich, 2014. For a comprehensive review see, Wolpe & Rowe, 2014).

Recent studies have started to look into the cognitive mechanisms and factors that underlie and modulate the intentional binding effect and more interestingly the SoA (cf., Cravo, Claessens, & Baldo, 2010; Ebert & Wegner, 2010; Engbert & Wohlschläger, 2007; Engbert, Wohlschläger, & Haggard, 2007; Hughes, Desantis, & Waszak, 2012; Moore, Lagnado, Deal, & Haggard, 2009). A multitude of models and theories ultimately converge into two central kinds: a predictive account and a retrospective account. These two forms describe two mechanisms that have been shown to take part in the formation of the intentional binding effect.

According to the predictive account, also referred to as internal forward model, two processes take place in the preparatory stage of action execution. First, a forward *dynamic* model describes a process that predicts the future state of the motor system and allows for error correction on the go, resulting in a smooth and adaptive motor action. Second, a forward *sensory* model describes a process that is responsible for the prediction of the sensory outcome of a specific action by learning the causal relations between actions and outcomes (Blakemore, Wolpert, & Frith, 2002; Miall & Wolpert, 1996; Wolpert & Ghahramni, 2000). Both processes take place during action selection and action execution but before the outcome of the action is known and can therefore generate a sense of agency prospectively.

While the predictive account puts the emphasis on the motor system, the retrospective inference account focuses on the processing of the sensory data. A retrospective inferential process is said to take place after a sensory outcome has been perceived and is responsible for determining the sources of actions and outcomes (cf., Aarts, Custers, & Wegner, 2005; and Wegner, 2003). In a framework called the comparator model, a comparison between the expected and the actual outcomes takes place after the outcome has been processed. If a comparison results in a match, a SoA is generated retrospectively, whereas a mismatch would lead to an attribution of the outcome to an external source (Frith, Blakemore, & Wolpert, 2000; recently reconsidered, Frith, 2012). Instead of relying on endogenous efferent copies of bodily actions, exogenous sensory cues (afferent feedback) are compared with expectations. While the comparator model stresses the importance of the intermediate processes of action generation, the post-dictive mechanism is also represented by the theory of apparent mental causation, formulated by Wegner and Wheatley (1999). The theory sets three prerequisites for the experience of agency: the priority, consistency, and exclusivity between intention (to act) and the action itself. An intention to act should temporally precede and be in proximity to an action. The intention should be compatible with the action. Finally, the intention should be the only apparent cause of the action. It is only when all three criteria are met that a retrospective sense-making process will generate a SoA (for more on both mechanisms see: Chamobon, Sidarus, & Haggard, 2014; Haggar & Chamobon, 2012).

For some time, predictive and inferential mechanisms have been regarded as competing and mutually exclusive explanations due to their different characteristics concerning their chronological timelines and epistemological bases. On the one hand, a predictive mechanism is assumed to take place before action execution and to rely on endogenous bodily signals (i.e., efference copies). On the other hand, a retrospective inference mechanism is considered to occur only after the brain has processed a sensory outcome and to rely on external stimuli. Although they have distinct characteristics, increasing evidence supports both mechanisms. This has led researchers to consider the integration of prediction and retrospective inference and to look for a more cohesive framework. One such leading theory, which makes use of the concept of Bayesian probability, is optimal cue integration (Moore, Wegner, & Haggard, 2009; Synofzik, Vosgerau, & Lindner, 2009; Synofzik, Vosgerau, & Voss, 2013; Wolpe, Haggard, Siebner, & Rowe, 2013). According to its central premise, action and outcome (i.e., the cues) contribute to their own shift in temporal perception (i.e., intentional binding) based on their relative reliability. A reliable action is to be understood here as one whose sensory information is precise. Such information comes from the action's ongoing efferent and afferent signals. Alternatively, a sensory outcome is more reliable when its signal to noise (S/N) ratio is high. Depending on a specific context, a weighted average is formed by a consideration of both cues, where the more reliable cue has a stronger effect.

Since the context of the action has a decisive influence on its perception and on the coupling process with a subsequent outcome, we were interested in investigating this relationship. In the current study we investigated prediction and inference and were interested in finding out how the long-term overall context of the action, as compared to its very recent preceding context, influences the action binding effect and which one better accounts for the two mechanisms. For that purpose, we employed a paradigm designed by Moore and Haggard (2008) to behaviorally dissociate the contribution of prediction and inference. Participants watched a rotating Libet clock and performed voluntary button presses, which were sometimes followed by a tone. First, we followed Moore and Haggard's probability manipulation: the probability of the tone within a block was manipulated to create an overall high or low prediction (predictive mechanism) in trials with or without tones (retrospective inference). This enabled a dissociation based on the long-term accumulation of action-outcome couplings. Second, we have adjusted Moore and Haggard's second analysis to fit our question: Originally, the analysis of the recent context of an action was carried out by including trials that were preceded by only one single action-tone trial. The amount of these preceding trials is therefore held constant and each trial is sorted according to one of three levels depending on the position of the action-tone trial (i.e., 100, 010, 001 where 1 represents action-tone trial and 0 represents action only trial). However, as we were interested in the accumulated amount rather than the distance of preceding trials, the second analysis was modulated as follows: we combined the two levels of probability (to cancel out their effect), and instead sorted each trial (with or without the tone, to find the inferential contribution) according to the amount of its recent preceding trials, depending on how many of them included action-tone couplings (more accumulated couplings stands for stronger prediction). Additionally, we have added a fourth level (i.e., "none" – no accumulation) to have a bigger amount of trials included in analysis as well as to be able to compare between the different accumulation levels and a no-accumulation level. The intentional binding effect of actions was calculated for these two types of analyses to reveal which better accounts for the effect and its underlying mechanisms.

Contingency and contiguity are both well-established factors known to shape the context that influences actions in instrumental learning (Shanks & Dickinson, 1991). While contingency has been shown to specifically modulate the perceived time of actions (Moore, Lagnado, et al., 2009), contiguity has not yet been directly tested with regard to the intentional binding effect (but see Moore & Haggard, 2008, partially supporting the role of contiguity in this context). We were inclined to favor the recent context based on the assumption that the proximity (i.e., the temporal contiguity) of the action-outcome couplings to the current trial would have a stronger influence compared to more distant accumulated trials. We hypothesized that the very recent accumulation of action-outcome couplings, rather than the overall probability of a block, would better account for both prediction and inference.

To further investigate this relation, we recorded brain activity with electroencephalography (EEG). As button press and tone operationalize action and outcome, we looked into the modulation of the corresponding event related potentials (ERPs): the readiness potential (RP) associated with the button press and the auditory evoked potential (AEP) associated with the tone. The RP is an ERP primarily related to the measurement of activity in the motor cortex and supplementary motor area of the brain leading up to voluntary muscle movement (Shibasaki & Hallet, 2006). A recent EEG study by Jo, Wittmann, Hinterberger, and Schmidt (2014) has brought evidence for the role of the RP as a neural correlate of the formation process of the intentional binding effect. Jo and colleagues reported that self-initiated movements that cause an outcome, following ongoing negative deflections in the early RP, result in a stronger binding effect compared to positive potentials of the RP. Specifically, the negativation of the early RP was found to be correlated with stronger backward shift of the outcome towards the action. So far, this is the only empirical evidence that directly supports the RP as a central neural correlate of the intentional binding effect. In the current study, we were interested to further study this relation and extend it to the mechanisms underlying the intentional binding effect of actions. Since the predictive mechanism is assumed to take place prior to action execution and to rely on endogenous sources, and taken together with the aforementioned previous findings, we argue that the RP is the neural marker best suited to study the contribution of a predictive mechanism to the binding effect of actions and the SoA. As in most other intentional binding studies, we used a tone as the sensory outcome of the action (e.g., Haggard & Clark, 2003). On the neural level, we therefore analyzed the modulation of the AEP. The AEP is another type of ERP that reflects the neural activity, which underlies the processing of auditory stimuli. AEPs can be analyzed to either express alterations in low-level perception (e.g., the N200 component) as well as to reflect higher cognitive processes (e.g., the P300 component), in which we were interested (Cone-Wesson & Wunderlich, 2003). As defined earlier, the retrospective inferential mechanisms is assumed to take place during the processing of the sensory outcome and to rely mainly on the outcome itself, rather than on internal bodily signals. Therefore, we expected the P300, the component of the AEP that reflects top-down cognitive processes during stimulus perception, to reflect the contribution of an inferential mechanism in the intentional binding effect. We hypothesized that both the RP and the P300 will reflect the intentional binding effect of actions through its modulation by either the long-term or the immediate context. The ERP analysis follows the behavioral results of the two analyses.

2. Materials and methods

2.1. Participants

Twenty-four undergraduate students from the Psychology Institute of the Humboldt-Universität zu Berlin (12 males, 12 females; mean age 24.7 years, SD = 4.9, range 18–37) took part in the present study. All participants were right-handed, had normal or corrected-to-normal vision and hearing, and had no history of psychiatric or neurological disorders. Participants were given course credit for their participation, signed an informed consent form and were given the right to quit participation at any point. The study was approved by the ethics committee of the Psychology Institute of the Humboldt-Universität zu Berlin and was conducted according to the declaration of Helsinki (WMA, version October 2013).

2.2. Apparatus and procedure

The experimental procedure was based on a paradigm by Moore and Haggard (2008) and is summarized in Fig. 1.

2.2.1. Task and experimental set-up

Participants were seated in front of a 17-in. computer screen on which a Libet-type clock was presented (for more on the original Libet method see, Libet, Gleason, Wright, & Pearl, 1983). The clock, 3.5° visual degrees in size, was rotating at a rate of 2560 ms per revolution (see Fig. 1(1)). To allow for a minimal “refractory period” between trials, every trial started with a fixation cross for 500 ms and was then followed by the clock. A clock-hand appeared in a random position and started rotating. Participants were asked to respond spontaneously by pressing the space bar at their own time with the index finger of the right hand. The button press was followed by a 250 ms interval after which a tone was played (1000 Hz for a duration of 100 ms). The clock-hand kept on rotating for a randomized period of time (between 1.5 and 2.5 s) and the trial ended with an estimation of either the button press or the tone. Participants typed in (with their left hand) a two-digit number that stands for the position of the clock-hand at the time of the event in question. When an estimation of the tone was not possible, a dummy number, ‘99’, was entered (see Fig. 1(2)).

2.2.2. Experimental design and manipulation

Each participant was presented with two types of baseline conditions (i.e., action-only and tone-only) and two types of experimental conditions (i.e., low tone probability and high tone probability). In the action-only baseline condition, participants pressed the button freely, and the clock kept on rotating for a random time. No tone followed. The time of the button press was then estimated

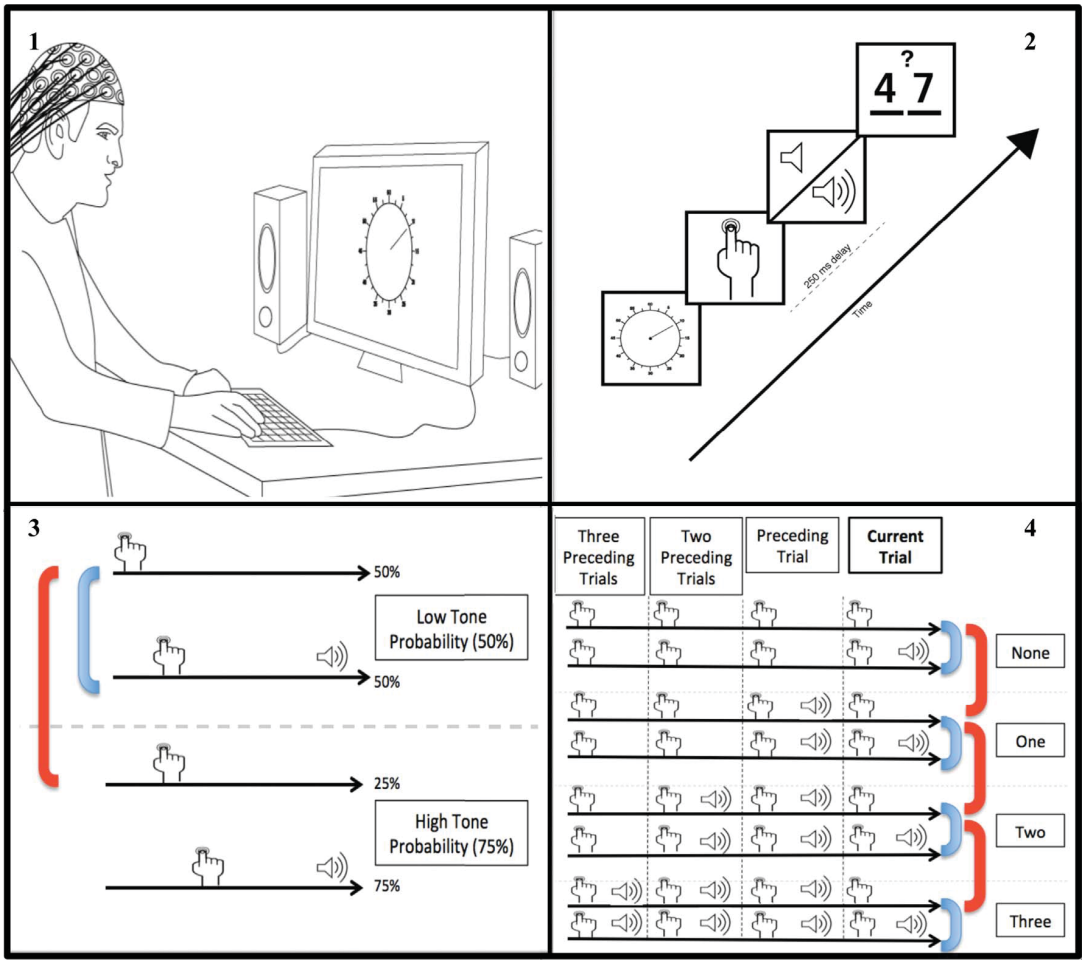

Fig. 1. (1) Experimental set-up. (2) Sequence of a single trial. (3) Long-term context of a single trial. Conditions contrasted to calculate the contribution of a predictive mechanism (red line) and an inferential mechanism (blue line). (4) Immediate recent context of a single trial. Multiple contrasts are averaged to estimate the contribution of a predictive mechanism (red lines) and an inferential mechanism (blue lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(based on the position of the clock-hand at the time of the event). In a tone-only baseline condition, participants were instructed not to press the button and only wait for the tone, which was played randomly between 2 and 6 s from the trial's onset. The time of the tone had to be estimated. The experimental conditions consisted of blocks with either 50% (low) or 75% (high) probability of trials with a tone following the button press. The estimation of either action or tone was blocked. In total, each participant had to complete 2 baseline blocks and 12 experimental blocks, 6 of each probability level divided to 3 blocks for every judged event. Each block consisted of 32 trials and in the experimental blocks the tone/no-tone trials were pseudo-randomly distributed. To control for learning, carry-over, fatigue and other order effect, two sequences were created in advance (probability blocks were interleaved) and counter-balanced by assigning each participant to one of the two pre-determined sequences. Then, the type of event judgment block (action/tone) and the position of the baseline blocks were randomized for each participant anew. At the beginning of the session participants completed two short training blocks, one of each probability level, which were not included in analysis. Each session lasted around two hours. Presentation of the clock and collection of the response data were performed by MATLAB version 8.4 R2014b (MathWorks, Germany) using the Psychtoolbox version PTB_Beta-2015-04-19_V3.0.12.

2.3. Behavioral data analysis

To measure temporal binding, action and tone are better considered separately. This way, the unique temporal shift of each component – action and outcome – can be revealed and a more informative account of the effect is achieved compared to a

calculation of the total effect (i.e., action and outcome taken together). Our analysis focused on the binding of actions. A measure for action binding is calculated by subtracting the mean judgment of actions in the action only baseline condition from the mean judgment of actions in experimental conditions. The size and direction of the temporal shift is taken to be the quantitative implicit measure of the extent to which the subject had the experience of agency. Since the Libet-clock paradigm is a sensitive measure that can fluctuate on the trial level, extreme values have to be carefully excluded to avoid the distortion of results (for more on the factors influencing the time judgments using the rotating clock see Pockett & Miller, 2007). The averaged means presented were therefore trimmed as follows: as a first step, the standard deviations were calculated for each participant over all trials and blocks of each condition. Second, trials with values greater or smaller than two standard deviations from the mean of the specific condition were discarded. Lastly, the trimmed means were calculated and averaged over all participants. The total number of rejected trials amounts to 266, which is 4.9% of all trials.

First Analysis (long-term accumulation): The judgments of action times were subject to repeated measures ANOVA with probability level (high vs. low) and trial type (tone vs. no-tone) as within-subject variables. To isolate the relative contribution of predictive and inferential mechanisms to the temporal shift in action perception, specific cells of the 2 × 2 factorial design had to be contrasted (see Fig. 1(3)). First, the contribution of the predictive mechanism is obtained by subtracting no-tone trials in the low probability condition from no-tone trials in the high probability condition. In the absence of a tone, any difference between the two stems from the difference in outcome probability and reflects the presumed contribution of a predictive mechanism. Second, the contribution of the inferential mechanism is obtained by subtracting no-tone trials in the low probability condition from tone trials in the low probability condition. By holding the outcome probability constant (on the lower level), any difference between the two can be attributed to the tone and thus reflects the presumed contribution of an inferential mechanism.

Second Analysis (recent accumulation): In the second analysis we classified single trials according to the accumulated amount of action-tone trials that preceded them. Recent amount of action-tone couplings was considered as 10% of a block's length (i.e., three preceding trials. cf., Moore & Haggard, 2008 in their analysis of the influence of the distance of a single action-tone trial). Each of these preceding trials was then registered, resulting in four different levels of classification (see Fig. 1(4)): None- three preceding action only trials; One- first immediate preceding action-tone trial, two action only trials before that; Two- two immediate preceding action-tone trials, one action only trial before that, and Three- three preceding action-tone trials. To have more trials for analysis and to cancel out the influence of tone probability, data from both probability levels were collapsed. The new division resulted in a new 2 × 4 factorial design with trial type (tone vs. no-tone) and amount of recent accumulation (None, One, Two, Three) as within-subject variables. To isolate the unique contribution of each mechanism the following contrasts were calculated: Prediction – on no-tone trials, low amount of recent accumulation is subtracted from the subsequent higher amount (i.e., Three-Two, Two-One, One-None). The differences are than averaged to yield a single value. This follows the rationale, according to which in the absence of a tone, any difference between each two levels stems from the difference in amount of recent accumulation, which reflects a predictive mechanism. Retrospective inference - no-tone trials with a given amount of recent accumulation are subtracted from tone trials with the same amount of recent accumulation (e.g., Three(tone)-Three(no-tone), Two(tone)-Two(no-tone), etc.). The differences are than averaged to yield a single value. By holding the amount of recent accumulation constant, any difference between the two conditions can be attributed to the tone and thus reflects the presumed contribution of an inferential mechanism.

2.4. Electrophysiological recordings

QRefa Acquisition Software, version 1.0 beta (Max Planck Institute for Human Cognitive and Brain Science, Leipzig, Germany) was used for EEG and electro-occulogram (EOG) recordings. Scalp EEG was recorded from 44 Ag/AgCl electrodes embedded in an elastic electrode cap (Easycap GmbH, Germany) according to the international 10/20 system. EOG was recorded from electrodes placed above and below the right eye and on the outer canthus of each eye to monitor vertical and horizontal eye movements. The ground electrode was mounted between the eyebrows. The EEG and EOG were amplified with the Refa system (Twente Medical Systems International B.V.) and digitized online at a rate of 500 Hz (anti-aliasing low pass filter of 135 Hz). Data was referenced online to the Cz and impedance was kept below 5 kΩ.

2.5. Electrophysiological data preprocessing

Data preprocessing was implemented using the EEGLAB version 13_5_4b toolbox (Delorme & Makeig, 2004). EEG data were re-referenced to the average mastoids (A1 and A2) and band-pass filtered between 0.1 and 45.0 Hz (zero-phase filter with −6 dB cutoff) in order to correct for slow drifts and muscle artifacts. Continuous EEG data was segmented into event-locked epochs ranging from 1.3 s before the button press to 0.5 s after with first 300 ms as baseline correction and 0.2 s before the tone to 1 s after with first 200 ms as baseline correction. Epochs affected by artifacts (± 100 μV) of any electrode except for ocular movements were excluded for further analysis. To remove eye movements and further myogenic and technical artifacts, we performed an Independent Component Analysis (ICA) based on the logistic infomax algorithm by Bell and Sejnowski (1995). Component activations and scalp maps were visually inspected and components reflecting blinks, eye movements, or muscle contraction were excluded from further analyses. The remaining EEG data was then averaged for each participant and condition within the above mentioned time range.

Table 1
Mean judgment error and shift from baseline of action times - first analysis.

Condition (Tone probability within a block)	Trial type	Mean (SD) judgment error (ms)	Mean (SD) shift from baseline (ms)
Baseline action	Action only	−118 (60)	
Low tone probability (50%)	Action only	−114 (37)	4 (52)
	Action → Tone	−102 (50)	16 (45)
High tone probability (75%)	Action only	−113 (46)	5 (43)
	Action → Tone	−99 (52)	19 (49)

2.6. ERP analysis

ERPs were calculated in the appropriate regions of interest (ROIs): RP was measured over six lateral electrodes surrounding Cz, where preconscious activation leading to voluntary action is measured (M1, SMA and pre-SMA): FC1, FC2, CP1, CP2, C3 and C4. As participants always pressed the button with their right hand, we were interested in calculating the signal on the contra-lateral side. To do that, activity on three electrodes of each side was averaged. Then, the averaged signal on the right was subtracted from that on the left (Adapted from the calculation of the LRP on go-nogo tasks that involve action on both sides. See, Eimer, 1998). The AEP was measured around the midline, primary auditory cortex and auditory association areas where the processing of auditory stimuli is most evidently reflected (Picton & Hillyard, 1974): Pz, Cz, Fz, C3, C4, T7, and T8. The amplitude of the RP and AEP in each condition and participant was quantified by calculating the mean signal in the above-mentioned epochs and baseline corrected.

3. Results

3.1. Behavioral data

3.1.1. First analysis

Table 1 shows the mean judgment errors of action times in each condition as well as the mean shifts from the baseline condition. Values on the middle right column are calculated by subtracting the estimated time from the actual time of action on every trial. These are then averaged. Values on the right-most column are calculated by subtracting means of experimental conditions from the baseline mean. (See Table 4 in the appendix for the equivalent data of tone time estimations. As no hypotheses were formed regarding tone judgments, no further analysis was performed.)

Corresponding to our hypothesis, positive action binding was found in all experimental conditions – the perceived time of actions in the experimental blocks was later than the perceived time of actions in the baseline blocks. However, while tone trials on low probability level were marginally significant and those on high probability level turned significant ($t(23) = 1.62$, $p = 0.059$; $t(23) = 1.84$, $p = 0.039$), action only trials were not significantly shifted from the baseline ($t(23) = 0.34$, $p = 0.365$; $t(23) = 0.36$, $p = 0.355$, all one-tailed according to the hypothesized positive direction of the temporal shift, Bonferroni-corrected).

To investigate the relative contribution of predictive and inferential mechanisms we performed a 2 × 2 repeated measures

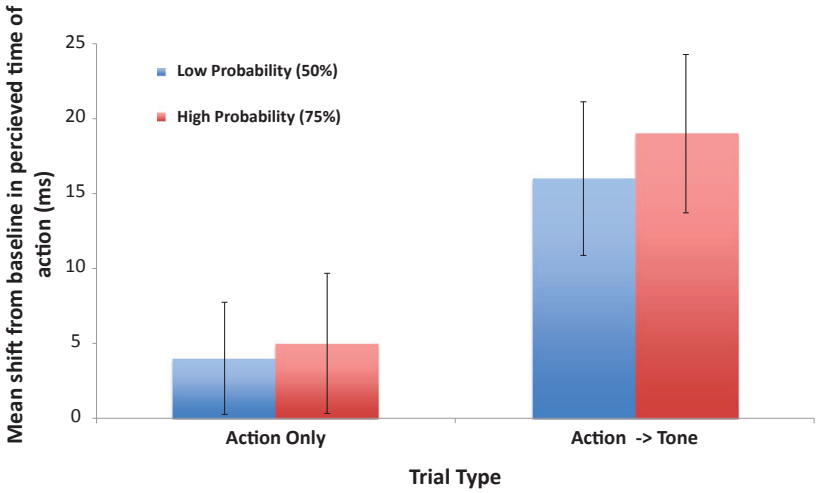

Fig. 2. Mean shift from baseline in the perceived time of action for every trial type and probability level. Bars represent the standard errors.

Table 2
Mean judgment error and shift from baseline of action times - second analysis.

Condition (Number of preceding action-tone trials)	Trial type	Mean (SD) judgment error (ms)	Mean (SD) shift from baseline (ms)
Baseline action	Action only	−118 (60)	
None	Action only	−140 (67)	−22 (66)
	Action → Tone	−109 (60)	9 (58)
One	Action only	−114 (64)	4 (63)
	Action → Tone	−112 (53)	6 (52)
Two	Action only	−110 (45)	8 (44)
	Action → Tone	−98 (43)	20 (49)
Three	Action only	−105 (51)	13 (50)
	Action → Tone	−92 (52)	26 (51)

ANOVA with probability level (low, high) and trial type (action only, action-tone) as within-subject factors. As could be expected from the primary results, there was no main effect of probability level ($F(1,23) = 0.077$, $p = 0.392$, $\eta^2 = 0.003$), but a significant main effect of trial type ($F(1,23) = 5.826$, $p = 0.012$, $\eta^2 = 0.202$) such that regardless of the level of probability, trials with tones showed a significantly stronger action binding than action only trials. There was no significant interaction between probability level and trial type ($F(1,23) = 0.167$, $p = 0.343$, $\eta^2 = 0.007$).

Fig. 2 shows the mean shifts from the baseline of the action time judgments for every experimental condition. As tone trials on the low probability level only approached significance and taken together with the fact that tone trials on the higher probability level and a main effect of trial type were found significant, we are inclined to cautiously conclude that an inferential mechanism is at work to some extent. However, this cannot be decisively determined. A predictive mechanism was not found present as action only trials were not significant on both probability levels and a main effect of probability level was not significant.

3.1.2. Second analysis

Table 2 shows the mean judgment errors of action times in each condition as well as the mean shifts from the baseline condition. The increase in action binding can be easily noticed and supports the positive association between the amount of recent action-tone coupling and the perceived forward shift of action time.

A repeated measures ANOVA with amount of recent accumulation (none, one, two, three) and trial type (action only, action-tone) as within-subject factors was performed. Analysis revealed a significant main effect of both trial type ($F(1,23) = 3.429$, $p = 0.038$, Greenhouse-Geisser, $\eta^2 = 0.130$) and amount of recent accumulation ($F(3,69) = 3.772$, $p = 0.007$, $\eta^2 = 0.141$) and no significant interaction ($F(3,69) = 1.096$, $p = 0.178$, $\eta^2 = 0.045$). The main effect of trial type shows that regardless of the level of recent accumulation, trials with tones showed a significantly stronger action binding than action only trials. The main effect of amount of recent accumulation shows that the action binding gets stronger the more action-tone trials precede a given trial, regardless of its type (with or without tone). Results allowed us to account for both mechanisms.

On no-tone trials, paired samples *t*-tests (one-tailed as action binding was assumed to be positively influenced by level of recent

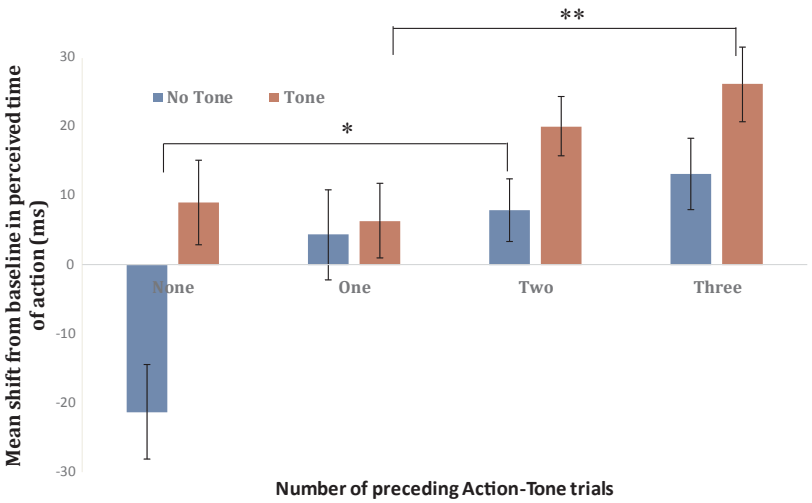

Fig. 3. Mean shift from baseline in the perceived time of action for every trial type and amount of recent accumulated action-tone trials. $^{*}p < 0.05$, $^{**}p < 0.01$. Bars represent the standard errors.

accumulation, Bonferroni-corrected) revealed significant differences between levels of recent accumulation none and one ($t(23) = -1.818$ $p = 0.041$), none and two ($t(23) = -1.924$ $p = 0.033$), and none and three ($t(23) = -2.178$ $p = 0.020$). On tone trials, paired samples *t*-tests revealed a significant difference between levels one and three of recent accumulation ($t(23) = -2.897$ $p = 0.004$). All other pairs were not significant or just marginally significant (no-tone trials: levels one-two $t(23) = -0.382$ $p = 0.353$, one-three $t(23) = -0.739$ $p = 0.233$, and two–three $t(23) = -0.663$ $p = 0.255$; tone trials: levels none-one $t(23) = 0.216$ $p = 0.415$, none-two $t(23) = -1.086$ $p = 0.144$, none-three $t(23) = -1.340$ $p = 0.096$, one-two $t(23) = -1.632$ $p = 0.058$, and two-three $t(23) = -0.826$ $p = 0.208$). Fig. 3 shows the mean shifts from the baseline of the action time judgments according to the second analysis. Additionally, to examine the single negative shift from the baseline, a paired samples *t*-test was conducted to compare between level none of recent accumulation on action only trials and the baseline and revealed no significant difference ($t(23) = -1.320$ $p = 0.199$, two-tailed).

3.2. Electrophysiological data

The amplitudes of the RPs on no-tone trials and the amplitudes of AEPs on tone trials were analyzed across different amounts of recent accumulation. In order to define the time windows for the analysis of the grand averaged RP and AEP (across participants) we performed an a priori repeated measures ANOVA: one-way ANOVA of RP with recent accumulated couplings and two-way ANOVA of AEP with electrode and recent accumulated couplings as within-subject factors. The ANOVAs were performed with a sliding window of 50 ms in consecutive 25 ms steps across the whole 1 s preceding the button press (i.e., −1000 to −950 ms; −975 to −925 ms; −950 to −900 ms, etc.) and the first half a second following the tone (i.e., 0–50 ms; 25–75 ms; 50–100 ms etc. – since the central components and main three segments of the AEP are all situated within the first half a second after the stimulus). Whenever more than three succeeding time windows revealed either a significant main effect of recent accumulation (in RP) or a significant interaction between electrode and level of recent accumulation (in AEP), these bigger time windows were further analyzed (Schaadt et al., 2015).

3.2.1. Readiness potential

The a priori ANOVA revealed significant effects in the consecutive time windows between −500 and −275 ms (see Table 5 in the appendix for a complete report of all time windows). A second-step one-way ANOVA that was conducted on this bigger time window revealed a significant effect of amount of recent accumulation ($F(1.917, 63.27) = 3.076$, $p = 0.034$, Greenhouse-Geisser, $\eta^2 = 0.219$). Post hoc pairwise comparisons (Bonferroni-corrected) revealed marginally significant differences between levels one and two of recent accumulated couplings (MD = 1.540, SE = 0.613, $p = 0.085$) as well as between levels one and three (MD = 1.277, SE = 0.501, $p = 0.081$). All other comparisons were not marginally significant. Fig. 4(1) shows the RP on the six electrodes across which analysis was performed as well as the topographic maps to illustrate the difference in activation between the levels of recent accumulation in the significant time window (Fig. 4(3)).

3.2.2. Auditory evoked potential

The a priori ANOVA revealed a significant main effect of recent accumulation and a significant interaction between electrode and level of recent accumulation in the consecutive time windows between 300 and 500 ms (see Table 6 in the appendix for a complete report of all time windows). A second-step ANOVA that was conducted on this bigger time window revealed a significant main effect of recent accumulation ($F(3,69) = 4.610$, $p = 0.004$, $\eta^2 = 0.295$) and a significant interaction between electrode and recent accumulation ($F(4.492,103.316) = 2.767$, $p = 0.016$, Greenhouse-Geisser, $\eta^2 = 0.201$). Post hoc pairwise comparisons (Bonferroni-corrected) showed a significant difference between levels none and two (MD = 1.748, SE = 0.447, $p = 0.007$) and between levels none and three (MD = 1.743, SE = 0.539, $p = 0.024$) of amount of recent accumulation. The significant interaction stems from differences on electrodes Cz, Fz, C3 and C4 (see Table 3).

Fig. 5(1) shows the AEP on the seven electrodes across which analysis was performed as well as the topographic maps to illustrate the difference in activation between the levels of amount of recent experience in the significant time window (Fig. 5(3)).

4. Discussion

In the current study we investigated the underlying mechanisms of the intentional binding effect, an implicit measure of the sense of agency. To uncover how prediction and inference are reflected in brain activity we combined behavioral and electrophysiological measures. We compared the effects of long-term and immediate context of a current trial on the binding effect of actions and were interested to find out which better accounts for its underlying mechanisms. We hypothesized that the amount of recent accumulated action-outcome coupling rather than the long-term context would better account for the two mechanisms. Furthermore, we recorded brain activity and analyzed the RP associated with the button press and the AEP associated with the tone. We hypothesized that RP and the P300 component of the AEP would reflect the significant modulations of the binding effect of actions, specifically the contribution of prediction and inference.

The study yielded the following main results: First, long-term accumulated action-outcome coupling partially accounted for an inferential mechanism, but we did not manage to replicate the predictive mechanism as reported by Moore and Haggard. Second, recent accumulated action-outcome coupling accounted for both the inferential and the predictive mechanisms. Third, the RP

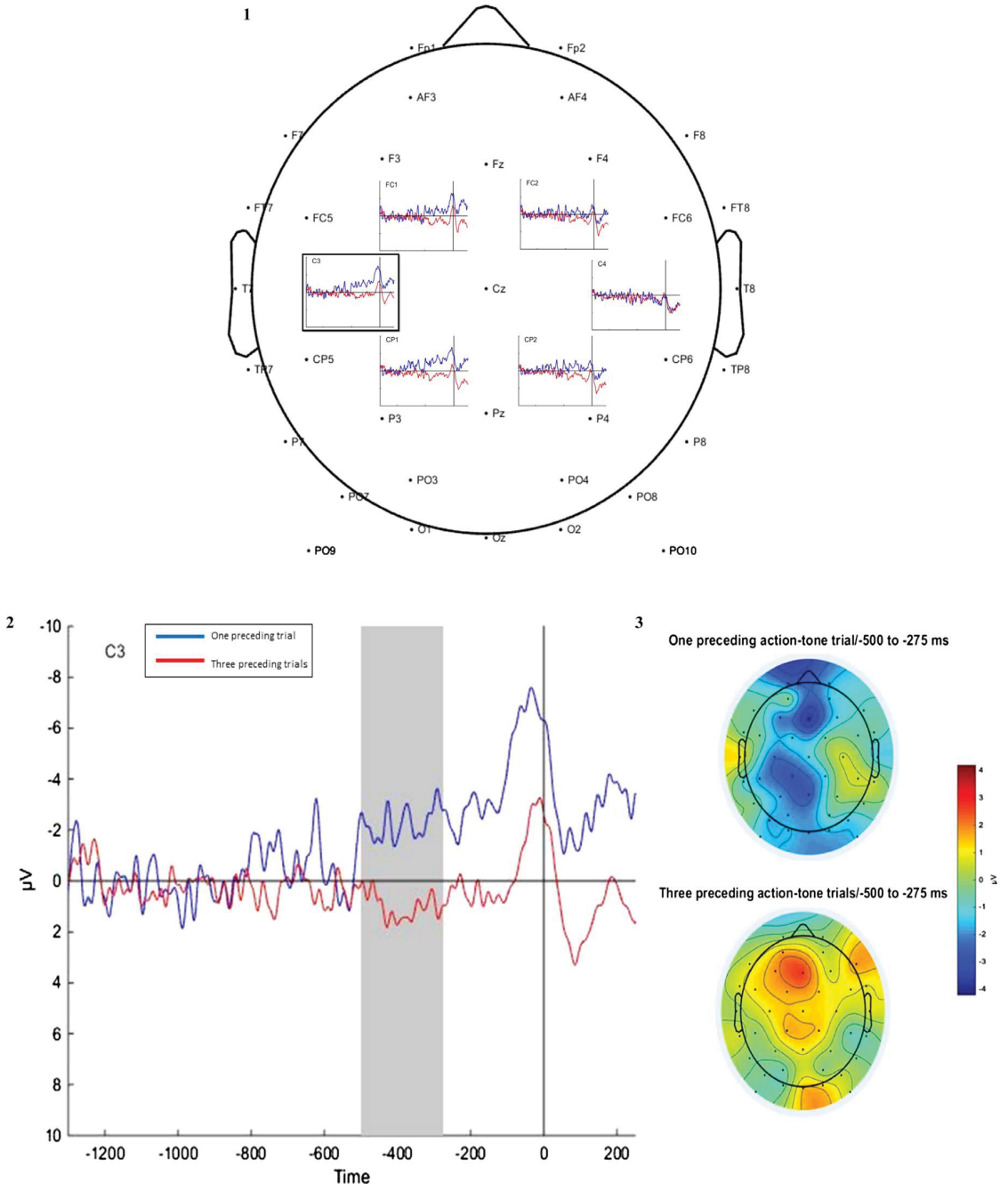

Fig. 4. (1) Grand averaged readiness potentials (RPs) as measured over six electrodes surrounding the motor cortex for no-tone trials preceded by one (blue) or three (red) action-tone trials. (2) Significant time window marked by the gray rectangle on electrode C3. 0 stands for the time of the button press. (3) Topography head maps showing the averaged spread of activation in the significant time window, compared between conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reflected the different levels of recent accumulated action-tone trials (and IB) by showing reduced negative amplitudes for trials that were preceded by more action-tone trials. Post hoc comparisons were not significant. Fourth, the AEP reflected the different levels of recent accumulated action-tone trials (and IB) by showing smaller positive amplitudes of the P300 component. The results will now be discussed and interpreted.

Table 3
Post hoc pairwise comparisons of levels none, two and three of recent accumulation across seven electrodes of the AEP between 300 and 500 ms after tone onset.

Electrode	Pairwise comparison	Mean difference (SE)	Significance (p)
Cz	None-Two	2.922 (0.782)	0.010
	None-Three	2.613 (0.906)	0.044
Pz	None-Two	1.331 (0.539)	0.093
	None-Three	1.532 (1.026)	0.491
Fz	None-Two	2.656 (0.875)	0.034
	None-Three	2.598 (0.805)	0.024
C3	None-Two	1.754 (0.647)	0.062
	None-Three	2.034 (0.645)	0.027
C4	None-Two	2.092 (0.498)	0.004
	None-Three	2.111 (0.745)	0.049
T7	None-Two	0.887 (0.663)	1.000
	None-Three	0.782 (0.556)	1.000
T8	None-Two	0.602 (0.479)	1.000
	None-Three	0.531 (0.366)	1.000

Behavioral results of the first analysis (i.e., long-term accumulation) could not replicate those of Moore and Haggard (2008) in that greater temporal shifts were found only for trials with tones compared to trials without tones, but not for higher compared to lower tone probability levels. When probability level is held constant, any difference in the temporal shift of the action is interpreted as the contribution of an inferential mechanism. Even though all experimental conditions showed a positive shift from the baseline, we did not find a significant main effect of probability level or a significant interaction between trial type and probability. The predictive component could not be accounted for by looking at the effect of overall probability.

It is interesting to note that our failure to replicate this result mirrors the pattern of results of schizophrenic patients in Voss et al., 2010. In their study, Voss and colleagues employed the same paradigm and compared between healthy subjects and schizophrenia patients. Their findings reveal a crossover effect, showing that action binding in healthy controls strongly relies on a predictive component while schizophrenia patients only show action binding as a result of an inferential process. A comparison between the results of healthy controls in Voss et al. (2010), Moore and Haggard (2008), and the current study, reveals three distinct patterns: predictive, combined and inferential -driven action binding, respectively. As all three experiments have employed an almost identical design, it is unclear what could account for the substantial differences. However, we would like to point out one crucial difference that can potentially underlie the fact that while both previous studies have supported a predictive component in healthy subjects, such contribution was found not significant in the current study. In both previous experiments, a prediction may have developed over time either through short but uninterrupted sequence of blocks (i.e., Voss et al. used a smaller number of blocks but excluded tone judgment blocks that shift attention from the action) or through a bigger number of blocks, interleaved with tone judgment blocks (i.e., Moore and Haggard). Due to the fact that in our study participants had to remain still to enable a clean EEG recording the total length of the experiment was confined and a smaller number of blocks per probability level was used. Additionally, we decided not to exclude tone judgment blocks in order to follow the original paradigm as closely as possible under the new circumstances and to control for a biased shift of attention towards the action. It might be that the combination of the number of blocks and the decision to include tone judgment blocks in the general sequence interrupted the overall development of a predictive mechanism. Further research is needed to shed light on the precise factors and environmental features that facilitate or hinder the generation of prediction in the coupling of actions to their sensory outcomes.

In a second analysis, we sorted each single trial according to the trials immediately preceding them. In this way, a continuum of four degrees of recent accumulated action-tone couplings was created for trials with and without tones. The average shift from the baseline in action time was calculated for every one of these four levels. Results showed a consistent positive relation between the amount of accumulated action-tone couplings and the perceived shift in action time: the more action-outcome trials preceded a single trial, the stronger was the temporal action binding. This was found to be true for no-tone and tone trials, a fact that suggests a common ground for the predictive and inferential components in the form of accumulated recent action-outcome couplings. One exception in this regard is the negative shift observed on no-tone trials preceded by three action only trials. However, since this shift was not found to be significantly different from the baseline, we do not discuss it here further. Significant main effects of trial type and amount of recent accumulation corroborated our hypothesis that the immediate context of a single trial rather than the overall long-term context accounts for both mechanisms. In comparison to the findings of the second analysis of recent experience by Moore and Haggard (2008), who found that “recent experience generates a predictive shift in the awareness of action on ‘action only’ trials, but not a significant change in inferential action awareness on ‘action + tone’ trials”, our findings suggest that both trial types are affected by the amount of recent accumulated action-tone trials.

To further investigate the relation between recent accumulation and the intentional binding effect we analyzed the accompanying brain activity at the time of the task. EEG activity was divided in two temporal sections, to account for activity before action execution and after stimulus presentation. Readiness potentials and auditory evoked potentials were epoched and averaged across subjects to reveal differences between conditions.

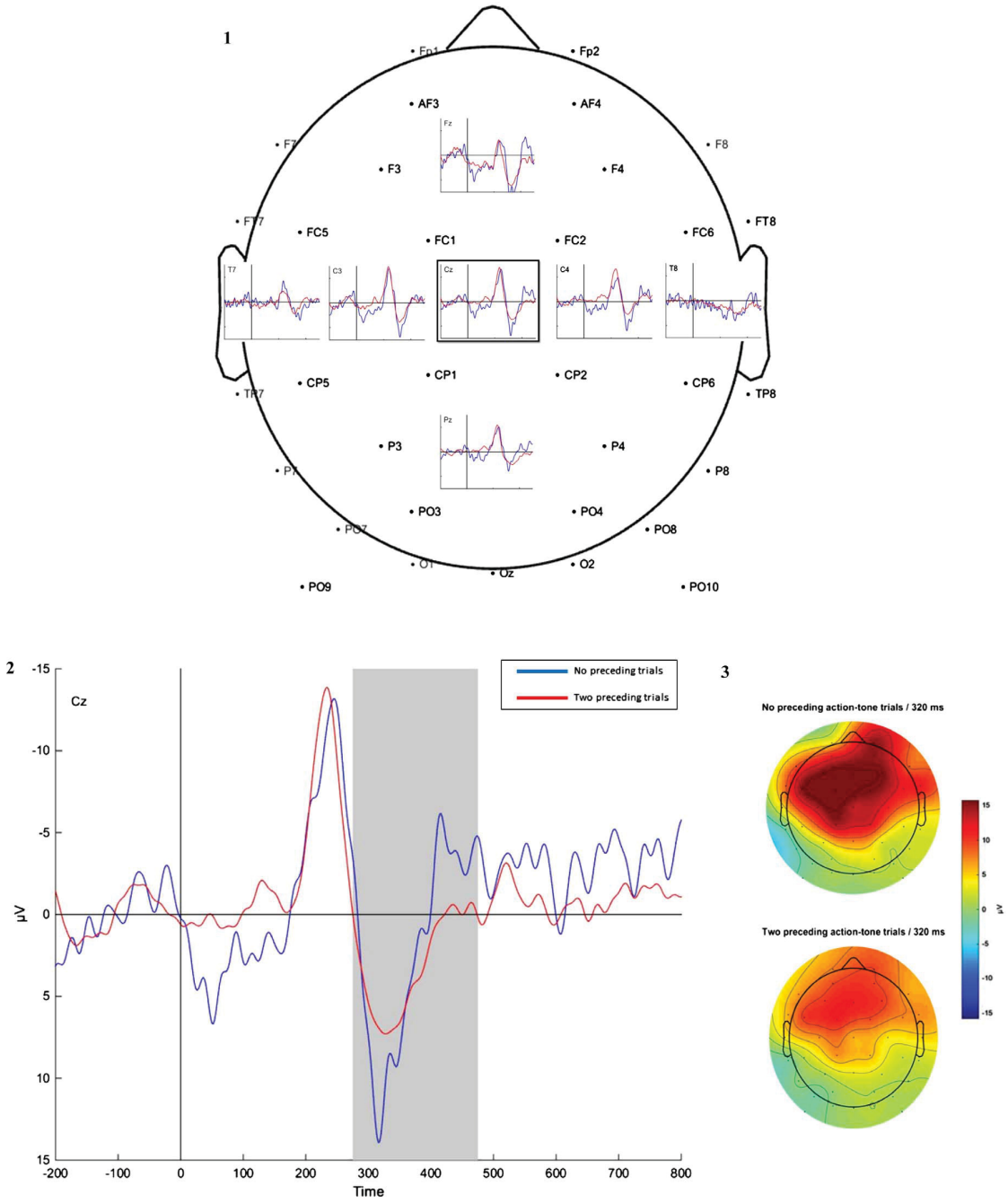

Fig. 5. (1) Grand averaged auditory evoked potentials (AEPs) as measured over seven electrodes for tone trials preceded by no (blue) or two (red) action-tone trials. (2) Significant time window marked by the gray rectangle on electrode Cz. 0 stands for the time of the tone. (3) Topography head maps showing the spread of activation at 320 ms (i.e., the P300), compared between conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

First, no-tone trials with more recent accumulated action-tone couplings were reflected in significantly reduced negative amplitudes of the RP compared to those with less accumulated couplings. Post hoc pairwise comparisons were conducted to investigate this modulation but did not reveal significant differences. In the absence of a tone, the difference in action binding effect and RP amplitudes is attributed to the different levels of recent accumulated action-tone couplings that preceded them. This is considered as supporting the expression of a predictive mechanism on the neural activity prior to action execution on trials where a stronger action

binding effect is present. However, as the post hoc results were only marginally significant and as the current study only considers the binding of actions towards outcomes, the results should not be simply generalized over the binding of outcomes. A study by Jo et al. (2014) showed that the early RP is significantly correlated with stronger outcome binding towards actions. The reported results, however, show that a stronger backward shift of outcomes is reflected in more negative amplitudes of the early RP as compared to a weaker binding. Considered in light of our findings that point to an opposite relation between RP amplitudes and action binding, it can be said that the relation between the RP and the temporal shifts of action and outcomes is to be considered cautiously. When action and outcome binding are considered separately, different and even opposite trends are revealed on the neural level. A more encompassing framework will have to be employed in order to explain the overarching relation between the perceived temporal shift and the underlying neural activity as is expressed by the RP. Especially beneficial would be a paradigm that can account for the binding of actions and outcomes simultaneously but enable their separate analyses, such that a single trial could be analyzed to reveal both components. Moreover, our analysis of the RP focuses on its later segment and leaves out the early tail. The decision to use this time range for the action-locked epochs stemmed from the fact that participants' actions were not restricted to a given point in the trial (unlike for example in Desantis, Hughes, & Waszak, 2012). We found it important that participants react spontaneously and without fixating on any specific cue. As a result, the button presses in some of the trials occurred very early, a fact, which hardened the analysis of the early tail of the RP. To overcome this drawback, future experiments would have to use bigger gaps between trials to enable a cleaner measurement of the whole trial.

Second, our findings point to the relation between a stronger temporal binding of actions and a diminished P300 component, which is modulated through higher levels of preceding action-outcome couplings. Tone trials displayed a significantly smaller positive amplitude around the P300 component of the auditory evoked potential for trials with more compared to less (or none) recent accumulated couplings. The P300 is traditionally divided into two separate sub components, the P3a and P3b. Judging by the latency of the peaks of the P300 (i.e., around 320 ms) and their temporo-parietal localization on the scalp, the component should be interpreted as a P3b (Linden, 2005). A central characterization of the P3b is its tendency to be elicited by improbable events. The less probable the event is the larger the P3b amplitude would be. Our results confirm this by showing larger positive amplitudes on tone conditions that were preceded by only one or none action-tone trials as compared to trials preceded by more action-tone trials. The alteration in the P3b is to be considered as part of a higher cognitive inferential process related to a shift in binding. It is important to note, that the alteration in the AEP at the P300 cannot be interpreted as reflecting a change in the low level sensory perception of the stimulus. Even though sensory attenuation is another indication of outcomes of self-generated movements, a much earlier segment of the AEP (e.g., N200 or even earlier) would have to show significant change (for more about sensory attenuation and its neural substrates see, Hughes & Waszak, 2011; Waszak, Cardoso-Leite, & Hughes, 2012). The contribution of an inferential component to the action binding effect is therefore reflected in the overall stronger binding in tone trials compared to no-tone trials across all levels of recent action-outcome couplings. On the neural level however, a direct contrast between tone and no-tone trials was not possible due to the fact that a comparison between AEPs had to include only trials with tones. Due to that shortcoming, a straightforward conclusion cannot be drawn and further investigation is needed in order to corroborate the relation formed here. Nonetheless, we are inclined to interpret the consistently stronger action binding in tone compared to no-tone trials on the behavioral level (which is additional to the action binding getting stronger through levels of accumulation) as related to an additional neural source (i.e., the cognitive process taking place around 300 ms post stimulus onset) as shown in the modulation of the AEP. This relation is here to be seen as reflecting a potential neural basis of an inferential mechanism, but due to the aforementioned limitation, further research is required to firmly establish the causal relation between the two and clarify its nature.

To conclude, we have shown that the immediate context of our actions has a decisive influence on our temporal perception of the action-outcome relation. Compared to the long-term accumulative context, the recent history of action-outcome coupling could account for both a predictive and an inferential shift of actions towards outcomes, as expected by the intentional binding effect. Further support stems from modulations of the RP and the P300 of the AEP, which reflect the implementation of these higher cognitive mechanisms on the neural level. Future research on the relation between action and outcome binding, prediction and inference and the neural activity – especially such that could overcome the drawbacks of the current design – is needed to further clarify the complex interaction that generates the implicit sense of agency and its underlying neural substrates.

Acknowledgments

We would like to thank the members of the Cognitive Psychology group at the Institute of Psychology of the Humboldt-Universität zu Berlin for their contribution to a fruitful discussion. A special thanks goes to Dr. Gesa Schaadt for her support in the analysis of the electrophysiological data. This work was supported by a doctoral grant of the Berlin School of Mind and Brain as part of the excellence initiative of the Deutsche Forschungsgemeinschaft (DFG) held by M.G.

Appendix A

See Tables 4–6.

Table 4
Mean judgment error and shift from baseline of tone times.

Condition (Tone probability within a block)	Trial type	Mean (SD) judgment error (ms)	Mean (SD) shift from baseline (ms)
Baseline tone	Tone only	−90(37)	
Low tone probability (50%)	Action → Tone	−89(91)	1(81)
High tone probability (75%)	Action → Tone	−93(96)	−3(84)

Table 5
Results of a priori sliding window analysis of RP amplitudes.

Time window in ms before button press	F test	p-value
−1000, −950	1.691	0.094
−975, −925	2.279	0.049
−950, −900	2.548	0.036
−925, −875	1.791	0.084
−900, −850	1.071	0.187
−875, −825	2.576	0.035
−850, 800	3.150	0.019
−825, −775	2.235	0.051
−800, −750	1.435	0.125
−775, −725	1.323	0.141
−750, −700	1.880	0.076
−725, −675	1.700	0.093
−700, −650	1.224	0.158
−675, −625	0.951	0.213
−650, −600	0.624	0.302
−625, −575	0.575	0.318
−600, −550	1.476	0.119
−575, −525	2.224	0.052
−550, −500	1.616	0.102
−525, −475	1.451	0.123
−500, −450	3.129	0.019
−475, −425	4.254	0.006
−450, −400	3.392	0.014
−425, −375	3.087	0.020
−400, −350	2.304	0.047
−375, −325	2.393	0.043
−350, −300	3.691	0.010
−325, −275	2.490	0.038
−300, −250	1.718	0.103
−275, −225	2.648	0.048
−250, −200	3.981	0.008
−225, −175	3.128	0.019
−200, −150	2.256	0.050
−175, −125	3.029	0.036
−150, −100	2.918	0.042
−125, −75	2.225	0.068
−100, −50	2.718	0.046
−75, −25	2.757	0.029
−50, 0	2.417	0.042

Table 6
Results of a priori sliding window analysis of AEP amplitudes.

Time window in ms after the tone	F test (amount of recent accumulation) (amount of recent accumulation * electrode)	p-value
0, 50	1.202	0.162
	0.411	0.423
25, 75	1.129	0.168
	0.608	0.337
50,100	1.094	0.176
	0.779	0.274
75,125	0.190	0.451
	0.704	0.300
100, 150	0.242	0.433
	0.701	0.298
125, 175	0.108	0.477
	0.815	0.265
150, 200	0.357	0.392
	1.047	0.198
175, 225	0.789	0.254
	1.499	0.100
200, 250	0.378	0.384
	1.428	0.111
225, 275	3.019	0.022
	2.633	0.019
250, 300	5.158	0.002
	3.254	0.009
275, 325	1.529	0.112
	1.322	0.137
300, 350	0.506	0.340
	2.001	0.043
325, 375	5.304	0.002
	3.490	0.003
350, 400	5.864	0.001
	2.773	0.015
375, 425	4.150	0.006
	2.295	0.038
400, 450	3.946	0.008
	2.195	0.042
425, 475	4.844	0.003
	2.579	0.022
450, 500	5.422	0.002
	2.756	0.018

References

Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition*, *14*, 439–458. http://dx.doi.org/10.1016/j.concog.2004.11.001.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. http://dx.doi.org/10.1162/neco.1995.7.6.1129.

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, *6*(6), 237–242. http://dx.doi.org/10.1016/S1364-6613(02)01907-1.

Chamobon, V., Sidarus, N., & Haggard, P. (2014). From action intentions to action effects: How does the sense of agency come about? *Frontiers in Human Neuroscience*, *8*, 320. http://dx.doi.org/10.3389/fnhum.2014.00320.

Cone-Wesson, B., & Wunderlich, B. (2003). Auditory evoked potentials from the cortex: Audiology applications. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *11*, 372–377. http://dx.doi.org/10.1097/00020840-200310000-00011.

Cravo, A. M., Claessens, P. M. E., & Baldo, M. V. C. (2010). The relation between action, predictability and temporal contiguity in temporal binding. *Acta Psychologica*, *136*, 157–166. http://dx.doi.org/10.1016/j.actpsy.2010.11.005.

David, N., Obhi, S., & Moore, J. W. (2015). Editorial: Sense of agency: Examining awareness of the acting self. *Frontiers in Human Neuroscience*, *9*, 310. http://dx.doi.org/10.3389/fnhum.2015.00310.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. http://dx.doi.org/10.1016/j.jneumeth.2003.10.009.

Desantis, A., Hughes, G., & Waszak, F. (2012). Intentional binding is driven by the mere presence of an action and not by motor prediction. *PLoS ONE*, *7*(1), e29557. http://dx.doi.org/10.1371/journal.pone.0029557.

Dewey, J. A., & Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing? *PLoS ONE*, *9*(10), e110118. http://dx.doi.org/10.1371/journal.pone.0110118.

Ebert, J. P., & Wegner, D. M. (2010). Time warp: Authorship shapes the perceived time of actions and events. *Consciousness and Cognition*, *19*, 481–489. http://dx.doi.org/10.1016/j.concog.2009.10.002.

Eimer, M. (1998). The lateralized readiness potential as an on-line measure of central response activation processes. *Behavior Research Methods, Instruments, & Computers. 30*(1), 146–156. http://dx.doi.org/10.3758/BF03209424.

Engbert, K., & Wohlschläger, A. (2007). Intentions and expectations in temporal binding. *Consciousness and Cognition*, *16*, 255–264. http://dx.doi.org/10.1016/j.concog.2006.09.010.

Engbert, K., Wohlschläger, A., & Haggard, P. (2007). Who is causing what? The sense of agency is relational and efferent-triggered. *Cognition*, *107*, 693–704. http://dx.

doi.org/10.1016/j.cognition.2007.07.021.
Frith, C. (2012). Explaining delusions of control: The comparator model 20 years on. *Consciousness and Cognition*, *21*, 52–54. http://dx.doi.org/10.1016/j.concog.2011.06.010.
Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society B*, *355*(1404), 1771–1788. http://dx.doi.org/10.1098/rstb.2000.0734.
Haggar, P., & Chamobon, V. (2012). Sense of agency. *Current Biology*, *22*(10), R390–R392. http://dx.doi.org/10.1016/j.cub.2012.02.040.
Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition*, *12*, 695–707. http://dx.doi.org/10.1016/S1053-8100(03)00052-7.
Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*, 382–385. http://dx.doi.org/10.1038/nn827.
Hughes, G., Desantis, A., & Waszak, F. (2012). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, *139*(1), 1133–1151. http://dx.doi.org/10.1037/a0028566.
Hughes, G., & Waszak, F. (2011). ERP correlates of action effect prediction and visual sensory attenuation in voluntary action. *NeuroImage*, *56*(3), 1632–1640. http://dx.doi.org/10.1016/j.neuroimage.2011.02.057.
Jo, H. G., Wittmann, M., Hinterberger, T., & Schmidt, S. (2014). The readiness potential reflects intentional binding. *Frontiers in Human Neuroscience*, *8*, 421. http://dx.doi.org/10.3389/fnhum.2014.00421.
Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) – The unconscious initiation of a freely voluntary act. *Brain*, *106*, 623–642. http://dx.doi.org/10.1093/brain/106.3.623.
Linden, D. E. J. (2005). The P300: Where in the brain is it produced and what does it tell us? *Neuroscientist*, *11*(6), 563–576. http://dx.doi.org/10.1177/1073858405280524.
Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*(8), 1265–1279. http://dx.doi.org/10.1016/S0893-6080(96)00035-4.
Moore, J. W., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, *17*, 136–144. http://dx.doi.org/10.1016/j.concog.2006.12.004.
Moore, J. M., Lagnado, D., Deal, D. C., & Haggard, P. (2009). Feelings of control: Contingency determines experience of action. *Cognition*, *110*, 279–283. http://dx.doi.org/10.1016/j.cognition.2008.11.006.
Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: A review. *Consciousness and Cognition*, *21*(1), 546–561. http://dx.doi.org/10.1016/j.concog.2011.12.002.
Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, *18*, 1056–1064. http://dx.doi.org/10.1016/j.concog.2009.05.004.
Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II: Effects of attention. *Electroencephalography and Clinical Neurophysiology*, *36*, 191–199. http://dx.doi.org/10.1016/0013-4694(74)90156-4.
Pockett, S., & Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition*, *16*(2), 241–254. http://dx.doi.org/10.1016/j.concog.2006.09.002.
Schaadt, G., Männel, C., van der Meer, E., Pannekamp, A., Oberecker, R., & Friederici, A. D. (2015). Present and past: Can writing abilities in school children be associated with their auditory discrimination capacities in infancy? *Research in Developmental Disabilities*, *47*, 318–333. http://dx.doi.org/10.1016/j.ridd.2015.10.002.
Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, *19*(4), 353–360. http://dx.doi.org/10.3758/BF03197139.
Shibasaki, H., & Hallet, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*, *117*, 2341–2356. http://dx.doi.org/10.1016/j.clinph.2006.04.025.
Synofzik, M., Vosgerau, G., & Lindner, A. (2009). Me or not me – An optimal integration of agency cues? *Consciousness and Cognition*, *18*, 1065–1068. http://dx.doi.org/10.1016/j.concog.2009.07.007.
Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: An interplay between prediction and postdiction. *Frontiers in Psychology*, *4*, 127. http://dx.doi.org/10.3389/fpsyg.2013.00127.
Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., & Haggard, P. (2010). Altered awareness of action in schizophrenia: A specific deficit in predicting action consequences. *Brain*, *133*, 3104–3112. http://dx.doi.org/10.1093/brain/awq152.
Waszak, F., Cardoso-Leite, P., & Hughes, G. (2012). Action effect anticipation: Neurophysiological basis and functional consequences. *Neuroscience and Behavioral Reviews*, *36*, 943–959. http://dx.doi.org/10.1016/j.neubiorev.2011.11.004.
Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences*, *7*(2), 65–69. http://dx.doi.org/10.1016/S1364-6613(03)00002-0.
Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, *54*(7), 480–492. http://dx.doi.org/10.1037/0003-066X.54.7.480.
Wolpe, N., Haggard, P., Siebner, H. R., & Rowe, J. B. (2013). Cue integration and the perception of action in intentional binding. *Experimental Brain Research*, *229*(3), 467–474. http://dx.doi.org/10.1007/s00221-013-3419-2.
Wolpe, N., & Rowe, J. B. (2014). Beyond the "urge to move": Objective measures for the study of agency in the post-Libet era. *Frontiers in Human Neuroscience*, *8*(450), http://dx.doi.org/10.3389/fnhum.2014.00450.
Wolpert, D. M., & Ghahramni, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*, 1212–1217. http://dx.doi.org/10.1038/81497.
World Medical Association (2013). World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, *310*(20), 2191–2194. http://dx.doi.org/10.1001/jama.2013.281053.

The Attribution of Agency to Non-Human Co-Actors in a Joint Task: A Driving Scenario

Michael Goldberg[1, 2], **Florian Koller**[2], **Niko Busch** [3, 4], **Elke van der Meer**[1, 2]

[1] Berlin School of Mind and Brain, Humboldt-Universität of Berlin, Luisenstraße 56, 10117 Berlin, Germany
[2] Institute of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany
[3] Institute of Psychology, University of Münster, Fliednerstraße 21, 48149 Münster, Germany
[4] Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Münster, Germany

Correspondence: Michael Goldberg, Institute of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany. E-mail: michael.goldberg@hu-berlin.de

Abstract

When cooperating in a joint task with another person, an extended agentic identity is formed. However, it has recently been shown that this extended agency was overturned with a non-human co-actor. In the current study, we have adapted the classical Libet paradigm and tested the attribution of agency in a more naturalistic and ecologically meaningful environment using a driving simulator. Participants used the simulator together with either a human or a computer co-actor. Similar to the classic Libet paradigm, participants judged the timing of their actions (pedal presses) and outcomes (tones). We measured the intentional binding effect, a subjective temporal shift of actions and outcomes, which is regarded as an implicit measure of the sense of agency. In line with previous studies, we found that action binding was extended to the human co-actor and not to the computer. However, our results show that tone binding was apparent for both co-actors. Moreover, task relevant feedback on who acted first had no influence on the implicit measure of agency. Our results suggest that action and outcome should be differentiated by their susceptibility to contextual cues, such as the co-actor. Additionally, task-relevant feedback is ineffective with regard to pre-reflective agency.

Keywords: Sense of Agency; Intentional Binding Effect; Joint Action; Human-Computer Interaction; Autonomous Systems

1. Introduction

For almost two decades, the sense of agency (SoA) has attracted a growing amount of scientific attention and is now studied in a variety of different contexts (for a recent comprehensive review, see Haggard, 2017). Most generally, the term refers to the subjective feeling of control over one's own bodily actions and through them the sensory outcomes in the external environment. To this day, the majority of studies in the field have focused on investigating the phenomenal nature of the SoA, shedding light on its different aspects, modulating factors and underlying mechanisms (see David et al., 2015 for a concise review about the current state of the research field; see Sperduti et al., 2011 for a meta-analysis of underlying brain structures). As a comprehensive picture of this relatively new construct began to form, a few studies have turned to examine the SoA in a more applied framework. Alongside clinical research investigating the relationship between different psychopathologies and the SoA (among others, see Maeda et al, 2011 and Voss et al., 2010 for an investigation of schizophrenia and Sperduti et al., 2013 for autism), a new branch of applied studies has developed. These studies extended the research of the SoA to the field of human-computer-interaction (HCI). By incorporating concepts like levels of automation (Berberian et al., 2012), input modalities (Coyle et al., 2012) and system feedback (Farrer et al., 2008) – all of which have important implications to our feeling of control – the interface between the two fields (i.e., SoA and HCI) was established. As technology becomes an indispensable part of our daily activities, the new course of emerging research is now all the more relevant, informing both system designers as well as scientific human agency research (see Limerick et al., 2014 for more on the link between HCI and SoA).

Another recent important advancement in experimental research into the SoA is the inquiry of agency in the context of a joint task. This intriguing step taken from the single actor to the multiple actor tasks opened up a new range of questions. Primarily, it is this social aspect that for the first time addressed the issue of how we experience agency when cooperating with another person towards a common goal rather than when acting on our own. So far, research conducted in this regard has uncovered a multitude of important insights relevant for co-action (see among others Dolk et al., 2014; Wohlschläger et al., 2003; Tsai et al., 2006), but findings seem to point to an interesting phenomenon when it comes to the SoA: when two people act together in a joint task, a new agentic identity is being formed. This new agentic identity, a 'we' rather than an 'I', is generated insofar as the implicit, pre-reflective agency is extended beyond one's own actions and self-produced outcomes. That is to say, when a participant acts together with another co-actor, some sort of unified agency emerges and extends beyond the single actor (Obhi and Hall, 2011a).

Among the different methods used to measure the SoA (e.g., the joint Simon effect, sensory attenuation and a variety of explicit measures), the intentional binding effect (IB) remains a central player. The effect, originally described by Haggard and colleagues (2002a,b), is a perceived temporal contraction of the time interval between voluntary actions and their sensory outcomes. Specifically, when voluntary actions are closely coupled with sensory outcomes, the actions are perceived later, and the outcomes earlier, as compared to when either actions or sensory events occur in isolation. Since temporal binding occurs for self-initiated but not for passive or triggered actions (cf., Buehner, 2012), it is considered to be an implicit marker of the

SoA. In most cases, the IB effect is studied by employing the pragmatic, straightforward paradigm of the Libet-clock. Participants usually watch a rotating clock on a screen and are required to estimate the times of their actions (key press) or their actions' outcomes (tone). While different versions exist for the estimation method (see Wolpe & Rowe, 2014), each having its own advantages and disadvantages, one central downside of the paradigm is its low ecological validity. In other words, the Libet-clock task, as it is most commonly employed, is highly restricted to the experimental setup. This is the case in so far that generalizations from the experimental actions and outcomes to real life actions and effects involve extended interpretation and speculation. Even though most studies using the Libet-clock benefit from its accuracy, consistency and ease of use, a leap to a more natural and meaningful environment, in which real actions and outcomes take place, remains somewhat problematic.

The combination of these two recent developments (i.e., HCI and the joint task context) along with the use of the intentional binding effect has enabled a more direct comparison between two conditions of a joint task. On the one hand, cooperation with a human co-actor and on the other hand cooperation with a non-human (e.g., computer, machine) co-actor. In their study from 2011(b), Obhi and Hall used the Libet-clock paradigm and measured the intentional binding effect, comparing a joint task performed with a human confederate to the same task performed with a computer co-actor. Participants had to press a button, followed by a tone (i.e., operant condition), and estimate the times of actions and outcomes by noting the clock time of the event. These estimations are compared to estimations on baseline conditions, where action is not followed by a tone and tone is presented without pressing the button. On each trial, either the genuine participant or the co-actor (confederate/computer) acted first and caused the tone. One of the central findings from Obhi and Hall (2011b) showed a clear-cut difference between the two conditions: when participants estimated the time of their actions and outcomes in the human-human condition, the action-effect interval was significantly shorter for operant compared to baseline conditions (pointing to the implicit SoA). Specifically, the onset of actions was found to be significantly delayed compared to the baseline, while tone onsets showed a significant backward shift from the baseline. Taken together, forward action binding and backward tone binding imply that the interval between action and outcome is subjectively shortened. Interestingly, binding effects were found for both the participant's own actions and outcomes, as well as when it is the co-actor's actions that produced the outcome. This is implying an extended SoA over the co-actors actions and outcomes, an implicit 'we' agency. However, when cooperating with a computer program instead of a human confederate, no binding effect (and so no implicit SoA) was found on any of the conditions. Moreover, not only binding was absent for action-effect intervals of the computer co-actor, but also binding for self-produced actions and outcomes was overturned under this context. That is, neither extended nor self-experience of agency were found.

The aim of the current project was to test the attribution of agency to human and non-human co-actors in a more ecologically valid environment and to overcome the theoretical leap between the highly restricted lab settings to the more noisy real world environment. Aspects of real life environments, involving unpredictability, changing regularities, or authentic context do not come into play in the classical paradigm. Specifically, our intention was to convert the original Libet-clock paradigm (Libet et

al., 1983) into a meaningful joint task, where participants are put in a more natural context. In order to do that, we used a driving simulator whereby participants are driving in a two-lane road side by side with another driver. The second driver was either a confederate, or a computer program, which was presented to participants as running the system of an autonomous car. The two-lane road was designed to converge at a certain point into a single lane. The goal of the joint task was to avoid an imminent crash at the convergence point by accelerating and overtaking the other driver in advance. In this new scenario, participants pressed a pedal (rather than a button) in order to accelerate the vehicle, which was followed by a tone signaling a successful avoidance of a crash. Participants performed the task twice with both the confederate and the computer, while unbeknown to them, the same behavior was employed by the program in both cases (that is, the confederate did not participate). The only difference between the two conditions was the belief manipulation about the partner with which the task is being performed. We shall first present our hypotheses and prediction. Then, in the methods section, a detailed description is given as to how the Libet-clock paradigm was adapted to suit the new scenario and what was used to replace each condition type (i.e., operant and baselines) to fit the driving scenario.

Based on the analysis by Obhi and Hall (2011b), three different predictions were made. The first two predictions (1a, 1b and 2) are in line with our intention to replicate Obhi and Hall's results, and thereby strengthening and extending the applicability of their findings to a naturalized scenario: 1a. Significant forward action binding would be found in the human-human, but not in the human-computer condition. 1b. Significant backward tone binding would be found in the human-human, but not in the human-computer condition. 2. Subjective action-effect intervals would be significantly shorter in operant compared to baseline conditions in the human-human, but not in the human-computer condition. The third prediction reflects our reasoning, according to which absence of self-agency found in the human-computer condition in the original study was due to the type of feedback given to participants. Obhi and Hall used color patches to signal whether the genuine participant or the co-actor acted first (and thereby caused the tone). One of three differently colored circles was presented directly after the tone and indicated that the first action was produced by 'self' (participant), 'other' (co-actor) or 'simultaneous' (both, later excluded from analysis). We argue that this method holds no meaningful relevance to the task in question. Although explicit agency beliefs (i.e., judgments) were formed by participants and reflected no doubt in the feedback, it had no effect on the implicit level of agency, as measured by the binding effect. In the current study, feedback was given on each trial following the outcome. After pressing the gas pedal to accelerate and hearing a tone, the overtaking was displayed, whereby one of the two cars overtook the other and moved forward to the converging road. We hypothesized that by generating a more seamless flow between the action, its effect and the feedback, we would achieve a stronger differentiation between self and other binding in the human-computer condition. In the original study by Obhi and Hall, the feedback factor was not significant. By contrast, we predicted that 3.the feedback factor would be significant only in the human-computer condition, such that for both actions and tones there would be a significant binding for self but not for the other's temporal estimations.

2. Materials and Methods

2.1 Participants

Forty-three undergraduate students from the Psychology Institute of the Humboldt-Universität zu Berlin (13 males, 30 females; mean age 26.2 years, SD=7.7, range 19-52) participated in the experiment. All participants were right-handed, had normal or corrected-to-normal vision and hearing, and held a valid driving license (mean of 7.6 years since its acquisition, SD=7.4, range 1-35). Participants were given course credit for their participation and gave written informed consent prior to the beginning of the experiment. The study was conducted according to the declaration of Helsinki (WMA, version October 2013).

2.2 Apparatus and Procedure

The experimental procedure followed that of Obhi and Hall (2011b) with several modifications described in the following section. The experiment was programmed and performed in Unity version 5.5.1 (Unity Technologies SF, US, 2009) and responses were registered using a wheel and pedal set (Logitech, Driving Force GT, E-X5C19). Visual information was displayed on a 27-in. DELL UltraSharp LED monitor, and sound was played through a connected headset.

2.2.1 Task and experimental set-up

Two identical set-ups were placed in the same room one next to the other (See Fig. 1). In order to avoid having any visual or auditory information about the co-actor, a dark curtain separated the two set-ups and participants were given earplugs. Participants sat one meter away from the monitor and their gaze was directed at the center of the right visual field of the screen to simulate the position of a driver, driving on the right lane. We ensured all participants could hold the wheel and reach the pedals comfortably, and that the sound was heard through the headphones while using the earplugs.

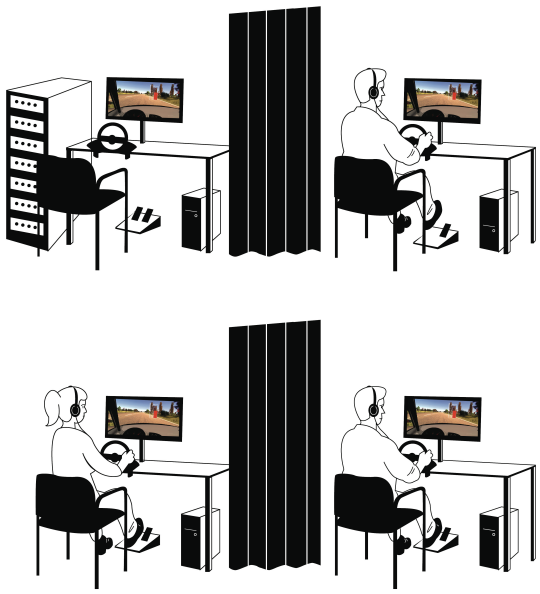

Figure 1: The experimental set-up. Two identical set-ups separated with a curtain. **(Top)** Human-computer condition. The participant seats next to the set-up on the right side. The other set-up remains empty and the server-like computer is turned on. **(Bottom)** Human-human condition. The participant seats on the right side, while the confederate seats on the left side. The server-like computer is removed from the room.

Each trial began with both cars positioned next to each other (participant's car on the right lane, the co-actor's car on the left lane) and a curvy road first had to be crossed. When reaching a straight part of the road, a street sign signaled the convergence of the two-lane road into a single lane (see Fig. 2). Once crossing the first street sign, a vertical bar (9 cm long) with 12 marks (ranging from 0 to 60, with steps of 5) appeared on the right side of the screen. Both the starting position of the bar (i.e., the amount to which it was filled) as well as its filling direction (i.e., up or down) were randomized each trial anew. The bar kept on filling up and down at a rate of 2.5 s in each direction, until reaching the second street sign (following Libet et al., 1983). Participants were asked to accelerate in order to overtake the other car and to avoid an imminent collision (i.e., the joint task). In order to achieve this task, participants were instructed to spontaneously (without fixating on a specific point) press the gas pedal once in the area between the first and the second street signs. The pedal press was followed by a 250 ms interval after which a tone was played (1000 Hz for a duration of 100 ms). The tone signaled the successful avoidance of a collision and the overtaking then proceeded automatically (i.e., displayed on the screen). The overtaking was pseudo-randomized across the block (with equal amount of trials for each feedback type: 'self' or 'other', depending on who acted first) and indicated to the participant who pressed the gas pedal first and caused the tone. Each trial ended with the required estimation of either the time of the pedal press or the tone. For the estimation, participants were presented with an identical empty bar and used a 24-position adjustment dial (placed on the wheel) to fill the bar to the estimated amount.

Figure 2: Operant trials. View from the participant's car during an operant trial (after crossing the first set of street signs). The vertical bar is seen on the right side and the second set of street signs can be spotted in the horizon.

2.2.2 Experimental design and manipulation

Participants completed the joint action task twice: once with another person – a confederate that was presented as a genuine participant – and a second time with a computer – allegedly run by a newly developed autonomous car system. For that purpose we have devised a cover story that enriched and strengthened the belief manipulation. Participants were told that cooperation between the cognitive psychology group of the Humboldt Universität zu Berlin and the AutoNOMOS Labs group (Freie Universität Berlin) was formed in order to test a new autonomous car system. Moreover, a bigger server-like computer was designed and stood next to the second (left) set-up. When the confederate was performing the task with the participant, the extra computer was shut down. When the participant was told he was performing the task with the computer, the extra computer was turned on and a simulation graphics was displayed on its screen. Both confederate and computer versions of the experiment were identical in all respects, with the only difference being the co-actor belief manipulation.

For each of the two partner types, two types of conditions, baseline and operant, were presented in separate blocks. In the action baseline condition, participants were driving on a straight road towards a highway (see Fig. 3 left). Once a participant crossed the highway road sign, she was requested to accelerate by pressing the gas pedal once, in a spontaneous fashion (avoiding a pre-planned press). No tone followed. The participant had to estimate the time of her pedal press. In the tone baseline condition, the participant was driving in a mountainous landscape (see Fig. 3 right). A road sign signaled that the driver is about to enter a hazard zone and was requested not to press the gas pedal from that point on in order to decelerate. A tone, signaling the end of the danger zone, was presented randomly in an interval of 2 to 6 seconds from crossing the road sign. The participant then estimated the time of the tone.
The operant, joint task conditions followed the scenario described in the previous subsection and were separated into different blocks, in which either the pedal press or the tone were estimated. As part of the cover story, participants were told that the tone would follow in a varying delay after the first of the two pedal presses. In reality, however, the tone always occurred 250 ms after the genuine participant's pedal press. In this manner, the confederate did not actively participate in the experiment.

Figure 3: Baseline trials. Before crossing the first set of street signs. **(Left)** Action baseline trials. **(Right)** Tone baseline trials.

In total, each participant completed two baseline blocks and two experimental blocks for each partner type, with each block consisting of 15 trials. To control for learning,

carry-over, fatigue and other order effects, four block sequences were created in advance and counter-balanced by assigning each participant to one of these four pre-determined sequences. At the beginning of the session, participants completed two free driving trials to get used to the simulator environment. Then, two training trials followed, one for each estimation type (i.e., pedal press and tone), which were not included in analysis. The order of confederate and computer co-actors was randomized between participants. Each session lasted around one and a half hours.

After the experiment, participants were questioned in a short debriefing session about the purpose of the study, whether they noticed anything about the procedures or had any thoughts about the other person or the computer. None of the participants showed serious suspicion with regard to the confederate, the computer or the cover story's general reliability, including the feedback given on each trial.

2.3 Data Analysis

The experiment comprised of a 2x2x2x2 multi-factorial, repeated measures design with the following factors: Co-Actor: Human, Computer; Condition Type: Baseline, Operant; Estimated Event: Pedal press, Tone; and Feedback: Self, Other. To corroborate our hypotheses we expected them to manifest in the results through an interaction including condition type and estimated event (for interval analysis only condition type) together with either co-actor or feedback. To enable a comprehensive and detailed interpretation of the data, a dual analysis was performed following that of Obhi and Hall (2011b). Shifts of single events (i.e., action and outcome) and changes in the derived action-outcome interval across conditions were calculated. The method incorporates elements of both the traditional treatment of data in intentional binding experiments (such as in Moore & Haggard, 2008, looking at action and tone separately) and the more recent studies that go beyond the strict definition of the IB effect, uncovering a broader range of effects by also referring to the interval measure (Strother et al., 2010).

A measure for action temporal binding is calculated by subtracting the mean judgment error of actions in the action baseline condition from the mean judgment error of actions in operant conditions. The same holds for tone (outcome) binding with the respective baseline and operant conditions. The size and direction of the temporal shift is taken to be the implicit quantitative measure of the extent to which the subject had the experience of agency. Our current design was based on the original Libet paradigm and followed closely the same parameters of the clock implemented in the vertical bar (i.e., rotation speed, rotation fashion and visual angle of the bar) and the estimation method, as well as all other decisive parameters from the intentional binding paradigm, including action-effect delay, tone characteristics etc.. Since the Libet-clock paradigm is known to be a sensitive measure that can fluctuate on the trial level, extreme values had to be carefully excluded to avoid the distortion of results (for more on the factors influencing the time judgments using the rotating clock see Pockett & Miller, 2007). The averaged means presented were therefore trimmed as follows: first, the standard deviations were calculated for each participant over all trials and blocks of each condition. Second, trials with values greater or smaller than three standard deviations from the mean of the specific condition were discarded together with values deviating by 500 ms from the time of the actual event. Finally, the trimmed means were calculated and averaged over all

participants. The total number of rejected trials amounts to 7.9% of all trials. The two analyses and their results are described in the following section.

3. Results

3.1 Single event analysis

To test hypotheses 1 and 3 a single event analysis was performed. As a first step, judgment errors were calculated on the trial level (i.e., distance of the estimated time from the actual time of the event). Then, trial values were averaged for each condition type and estimated event across all participants. This procedure was performed for both co-actors and feedback types. To test the hypotheses and investigate the differences in the temporal estimations of onset of single events, we ran a 2 (Co-Actor) by 2 (Condition Type) by 2 (Estimated Event) by 2 (Feedback) repeated measures ANOVA.

The analysis revealed several significant results: a main effect of condition type (operant conditions were always significantly shifted from baseline conditions: $F(1,42)=27.1$, $p<.000$), a main effect of estimated event ($F(1,42)=8.564$, $p=.006$) and an interaction between the two (stronger binding for tones than for actions: $F(1,42)=67.1$, $p<.000$). More interestingly, and relevant to our predictions, a significant interaction was found between co-actor, estimated event and condition type ($F(1,42)=6.141$, $p=.017$). Lastly, only marginally significant interaction was found between co-actor, condition type and feedback ($F(1,42)=4.053$, $p=.051$). All other main effects and interactions were not significant.

To investigate the 3-way interaction between co-actor, estimated event and condition type, paired samples t-tests were conducted (in order to exclude the feedback factor from the current analysis, data were collapsed across self- and other feedback trials; Bonferroni corrected). When participants performed the joint task with the confederate, estimates of the onset of the action and tone were both significantly shifted in operant compared to baseline conditions (action forward shift: $t(42)=3.026$, $p=.004$; tone backward shift: $t(42)=-10.3$, $p<.000$). In contrast, when participants performed the joint task with the computer's system, a significant temporal shift was found for the tone, but not for the action (action forward shift: $t(42)=1.568$, $p=.124$; tone backward shift: $t(42)=-7.292$, $p<.000$). Figure 4 shows the mean temporal estimations of action and tone on baseline and operant conditions for the human and computer co-actors. To summarize, intentional binding was found in the human-human condition for both action and tone. While action binding was not significant in the human–computer condition, tone binding was significant.

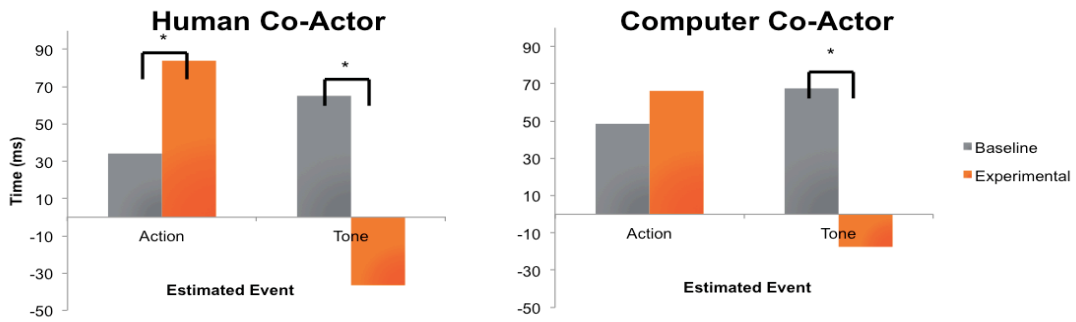

Figure 4: Single event analysis: mean temporal estimations. Mean temporal estimations of action and tone on baseline and operant conditions for the human **(Left)** and computer **(Right)** co-actors.

3.2 Interval analysis

To test hypothesis 2 an interval analysis was performed. To investigate the differences between baseline and operant action-effect intervals, four intervals were first calculated. The intervals were matched by feedback (self/other) and condition type (baseline/operant): 1. Action-baseline-self – tone-baseline-self; 2. action-baseline-other – tone-baseline-other; 3. action-operant-self – tone-operant-self; 4. action-operant-other – tone-operant-other. Means were then calculated over all participants for each of the four intervals. Finally, a 2 (Co-Actor) by 2 (Condition Type) by 2 (Feedback) repeated measures ANOVA was performed. Since Estimated Event was on this analysis already included in the calculation of the intervals, it was not added to the ANOVA as a separate factor.

The analysis revealed a significant main effect of condition type ($F(1,42)=67.16$, $p<.000$) and a significant interaction between co-actor and condition type ($F(1,42)=6.14$, $p=.017$). All other main effects and interactions were not significant. Paired samples t-tests (Bonferroni corrected) were conducted to investigate the source of the interaction and showed that the action-effect intervals on both the human ($t(42)=-7.630$, $p<.000$) and the computer ($t(42)=-6.403$, $p<.000$) co-actors were significantly shorter on the operant compared to the baseline conditions. In other words, intentional binding was found on the interval level for both types of co-actors. However, a further post-hoc t-test revealed a significant difference in the size of the binding effects between the two co-actors, such that the action-effect interval shortening is significantly bigger in the human compared to the computer co-actor condition ($t(42)=-2.527$, $p=.015$). These results cohere with the results from the single event analysis, insofar as they support a strong binding effect for the human co-actor (of both action and tone) and a weaker binding effect for the computer co-actor (of only the tone). Figure 5 shows the calculated intervals on baseline and operant conditions, for human and computer co-actors, as well as the total interval shifts.

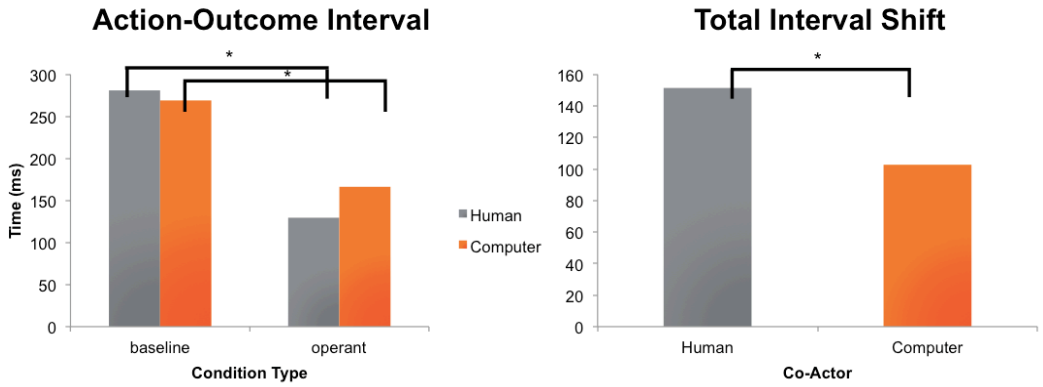

Figure 5: Interval analysis: mean derived intervals and total shifts. Derived action-outcome intervals **(Left)** on baseline and operant conditions, for human and computer co-actors, and the total interval shifts **(Right)** for both co-actors.

4. Discussion

In the current study, we investigated the influence of the type of co-actor in a joint task on the attribution of agency (i.e., its extension) for actions and their outcomes. Moreover, we were interested to find out whether task related feedback about the source of the action would be reflected in the implicit measure of a pre-reflective sense of agency. To that purpose, we adapted the classical Libet-clock paradigm and followed the belief manipulation designed by Obhi and Hall (2011b). These were embedded within a new driving scenario implemented by a 2-D driving simulator. Participants drove side by side with another person (confederate) or with a computer (autonomous car) and performed in a joint task. Temporal estimations of self-produced actions and outcomes as well as of actions and outcomes that were preceded by the co-actor were collected. Data was analyzed on both the single event and on the action-outcome interval levels.

The study yielded the following main results: First, a significant forward action binding was found in the human-human, but not in the human-computer condition. This finding corroborates our first prediction. Second, a significant backward tone binding was found on both co-actor conditions. This result fails to replicate the result from the original study by Obhi and Hall and contradicts our prediction that no significant shift would occur in the human-computer condition. Third, action-effect intervals were significantly shorter in operant compared to baseline conditions on both the human-human and the human-computer conditions. However, a further post-hoc test revealed a significant difference between the two interval shortenings such that a stronger binding effect was found on the human-human compared to the human-computer condition. This result contradicts our prediction, yet partially supports a difference between the two types of co-actors. Lastly, similar to the original study, and although adapted to the task's context, we could not find any effect related to the type of feedback (self or other) given to participants on any of the conditions. These results will now be discussed and interpreted.

Results of the single event analysis replicated those of Obhi and Hall by showing a significant forward action binding when cooperating with the confederate but not with the computer. This is regardless of whose action it was, which was indicated as the effective action. This result supports the central thesis that an extended agentic identity is formed when cooperating with another person in a joint task. Not only was binding found for one's own effective actions, but also when it was the other person's actions, which took place first and caused the outcome. When cooperating with the computer (the autonomous car), no action binding was found for self and other actions. That is, pre-reflective agency, as measured by the intentional binding effect, was neither extended to the computer, nor was it present for self-produced effective actions. Extended agency can be explained by appealing to the mirror-neuron system, responsible for the awareness and representation of another person's actions and intentions (Rizzolatti & Craighero, 2004; Iacoboni et al., 2005). Simply watching another person's action explains the activation of the same motor areas that would have been activated in the execution of the action, it is all the more expected to underlie the co-representation of a co-actor's actions in a joint task (see Sebanz et al., 2006 for an elaborated review of joint action mechanisms). Following this line of explanation, it is the lack of similarity between human and computer intentional and behavioral systems that is responsible for the inhibition of the attribution of agency to

the computer co-actor. Especially intriguing in this regard, was the failure to show binding for self-produced effective actions in the human-computer condition. One possible explanation for this finding is that participants' sense of efficacy is lost altogether, once cooperating with an automated system which might be assumed to employ a predetermined set of behaviors. Computer, and automated systems in general are regarded as fully programmed machines, lacking the ability to generate true intentions and decisions. Therefore, a different theory of mind is formed when it comes to understanding an automated co-actor's behavior. In a sense, it is as if the deterministic and unfamiliar nature of the computer's behavior takes over the ability to infer and form regularities about the action-effect pairings. Indeed, it has been shown that pre-reflective agency is generated by both predictive and retrospective mechanisms (Synofzik et al., 2013). By relying on external cues, sensory evidence is used to infer and make sense of the sources of actions and their outcomes. According to Wegner and Wheatley's theory of apparent mental causation (Wegner & Wheatley, 1999), this inferential process is dependent on three prerequisites: a thought must precede the action, must be consistent with the action, and any alternative source for the action must be excluded. Since a clear picture of the inner workings of the computer's system might remain unclear, it can be speculated that some or all of these preconditions for inference of self-agency are not met. However, further support and clarification is needed and might come from a design, combining different types of actors (for example, automated systems with different levels of humanized features) while also dissociating the unique contribution of the predictive (i.e., forward prediction models) and inferential mechanisms to the formation of the sense of agency.

The second part of the single event analysis revealed a backward shift of the tone for both co-actors. This result partially failed to replicate the finding by Obhi and Hall, since a backward tone shift was not expected on the human-computer condition. One major difference between the current and the original study was the newly designed context in which the joint task was performed. We are therefore inclined to see the diverging findings as reflecting this difference in the overall settings. Specifically, in the current study, the joint task departed from the nominal partnerships formed by merely watching the Libet clock. Participants were not simply involved in a symbolic task, but rather acted in order to achieve a common goal with overt consequences: when not acting in a timely manner, participants had to repeat the trial. When acting as expected, an overtaking of either one of the two cars took place. Moreover, the tones, originally having no conceptual relation to the button presses, were now playing a significant role by carrying task relevant meaning. We argue that it was these contextual differences that account for the extended binding effect not only for the human confederate but also for the automated computer.
One important question that has to be tackled at this point concerns the difference between actions and outcomes on the human-computer condition. What might underlie the difference between the two? We suggest that while the type of co-actor modulates the temporal estimations of actions for the above-mentioned reasons, this is not the case for temporal estimations of outcomes. The outcomes are independent from any higher-level knowledge concerning the co-actor, which is required for forming expectations or inferring about the event in question. In other words, while actions are tightly connected to the executing co-actor by being the direct result of a complex internal process of action initiation and preparation, their outcomes are external events only transitively connected to the agent (for more on internal forward

models and the comparator model see Synofzik et al., 2008). Hence, by participating in a joint task, the participant develops a non-discriminative approach for the desired outcome, as long as such outcome appears to settle with her existing expectations of the action-effect regularities.

Further support comes from a second analysis, in which we have looked into the differences between action-effect intervals. The analysis revealed that for both co-actors, intervals were significantly shorter in operant compared to baseline conditions, while a further post-hoc test highlighted that the interval shortening was significantly stronger on the human-human compared to the human-computer condition. This result lends support to our previous findings, where on the one hand strong tone binding effects were found for both co-actors (sufficient for the intervals to turn significantly shorter), and on the other hand, action binding was only apparent on the human-human condition (which is reflected in the post-hoc test). It is interesting to note that tone binding was found to be stronger than action binding regardless of the type of co-actor (shown by the significant effect of estimated event), strengthening our hypothesis that, although closely related, action and tone rely on different types of external cues, one of which being the type of co-actor.

Finally, explicit task relevant feedback had no significant effect on neither of the estimated events for both co-actors. In other words, while the participants' belief was shaped by the explicit feedback (shown to be accepted as reliable), pre-reflective agency seemed to be unaffected. This finding contradicts our prediction that task relevant feedback would show to be more effective than neutral feedback. However, the result corresponds with the established distinction between the 'feeling' and the 'judgment' of agency: "At the implicit level, a basic low-level feeling of being an agent is formed … This level is non-conceptual, and does not involve explicit agency attributions. Rather, experiences of action are simply tagged as self-caused or not … At the explicit level, a higher-order conceptual judgment of being an agent is formed … At this level, explicit attributions of agency to oneself or another are made." (Moore et al., 2012; For more on this distinction, cf.: Dewey & Knoblich, 2014). While we do not argue against the dissociation of the two levels of agency, we take the marginally significant interaction between co-actor, condition type and feedback, to point to the possibility that the form, type and temporal contiguity of feedback might have a varying influence not only on explicit judgments but eventually also on the implicit sense of agency. This assumption cannot be supported by our current findings but requires a direct comparison between several kinds of feedback. Future research might uncover whether some distinctive features (e.g., saliency, valence and so on) exist, such that these features can break the conceptual wall between the two facets of agency.

In conclusion, the current study lent external validity to former research findings on the attribution of agency to human and non-human co-actors. By embedding a joint task within a rich and meaningful driving environment, we have shown that action binding is present when cooperating with a human confederate but not with an automated system. The type of co-actor did not influence tone binding. We argue that in real life situations, action and outcome rely on distinctive external cues. Moreover, our results gave further support to the distinction between the explicit and the implicit aspects of agency, as feedback about the action source did not modulate the binding effect on any of the experimental conditions. Future research is needed in order to

deepen our understanding of the complex ways in which our sense of agency is formed in everyday interactions. Such progress will contribute to both the cognitive research of human agency as well as to future platform and interface designs, made to enhance and facilitate our interactions with automated systems. In a follow-up study, we are going to take a step in this direction by testing whether a human trained system and an anthropomorphized appearance can facilitate the implicit attribution of agency.

5. Acknowledgments

We would like to thank the members of the Cognitive Psychology group at the Institute of Psychology of the Humboldt-Universität zu Berlin for their contribution to a fruitful discussion (and Benjamin Schlotter for programming). We would also like to thank Prof. Dr. Raúl Rojas, supervisor of the AutoNOMOS Labs group (Freie Universität Berlin), for permission to use the group's name and logo as part of our cover story. This work was supported by a doctoral grant of the Berlin School of Mind and Brain as part of the excellence initiative of the Deutsche Forschungsgemeinschaft (DFG) held by M.G. as well as by the BIGS² research school at Humboldt-Universität zu Berlin (F.K.).

6. References

Berberian, B., Sarrazin, J-C., Le Blaye, P., and Haggard, P. (2012). Automation Technology and Sense of Control: A Window on Human Agency. *PLoS ONE*, 7(3): e34075, doi:10.1371/journal.pone.0034075

Buehner, M.J. (2012). Understanding the Past, Predicting the Future: Causation, Not Intentional Action, Is the Root of Temporal Binding. *Psychological Science*, 23(12), 1490-1497, doi: 10.1177/0956797612444612

Coyle, D., Moore, J., Kristensson., P.O., Fletcher, P.C., and Blackwell, A.F. (2012). I did that! Measuring Users' Experience of Agency in their own Actions. In *CHI, ACM Conference on Human Factors in Computing Systems*, (Austin, Texas, USA), 2025–2034

David, N., Obhi, S., and Moore, J.W. (2015). Editorial: Sense of agency: examining awareness of the acting self. *Frontiers in Human Neuroscience*, 9(310), doi: 10.3389/fnhum.2015.00310

Dewey, J.A., and Knoblich, G. (2014). Do Implicit and Explicit Measures of the Sense of Agency Measure the Same Thing? *PLoS ONE*, 9(10), doi: 10.1371/ journal.pone.0110118

Dolk, T., Hommel, B., Colzato, L.S., Schütz-Bosbach, S., Prinz, W., and Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. *Frontiers in Psychology*, 5(974), doi: 10.3389/fpsyg.2014.00974

Farrer, C., Bouchereau, M., Jeannerod, M., and Franck , N. (2008). Effect of distorted visual feedback on the sense of agency. *Behavioral Neurology,* 19, 53–57, doi: 10.1155/2008/425267

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18, 197-208, doi:10.1038/nrn.2017.14

Haggard, P., Aschersleben, G., Gehrke, J., and Prinz, W. (2002a). Action, binding and awareness. In Common mechanisms in perception and action. In W. Prinz and B. Hommel (Eds.). *Attention and performance*, XIX, 266–285. Oxford, UK: Oxford University Press.

Haggard, P., Clark, S., and Kalogeras, J. (2002b). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4), 382-385, doi: 10.1038/nn827

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., and Rizzolatti, G. (2005). Grasping the Intentions of Others with One's Own Mirror Neuron System. *PLoS BIOLOGY*, 3(3), 529-535, doi: 10.1371/journal.pbio.0030079

Libet, B., Gleason, C.A., Wright, E.W., and Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act.. *Brain*, 106, 623-642.

Limerick, H., Coyle1, D., and Moore, J.W. (2014). The experience of agency in human-computer interactions: a review. *Frontiers in Human Neuroscience*, 8(643), doi: 10.3389/fnhum.2014.00643

Maeda, T., Kato, M., Muramatsu, T., Iwashita, S., Mimura, M., and Kashima, H. (2011). Aberrant sense of agency in patients with schizophrenia: Forward and backward over-attribution of temporal causality during intentional action. *Psychiatry Research*, 198, 1-6, doi: 10.1016/j.psychres.2011.10.021

Moore, J., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, 17, 136–144, doi:10.1016/j.concog.2006.12.004

Moore, J.W., Middleton, D., Haggard, P., and Fletcher, P.C. (2012). Exploring implicit and explicit aspects of sense of agency. *Consciousness and Cognition*, 21(4), 1748-1753, doi: 10.1016/j.concog.2012.10.005

Obhi, S.S., and Hall, P. (2011a). Sense of agency and intentional binding in joint action. *Experimental Brain Research*, 211, 655–662, doi: 10.1007/s00221-011-2675-2

Obhi, S.S., and Hall, P. (2011b). Sense of agency in joint action: influence of human and computer co-actors. *Experimental Brain Research*, 211, 663–670, doi: 10.1007/s00221-011-2662-7

Pockett, S., and Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition*, 16(2), 241-254, doi: 10.1016/j.concog.2006.09.002

Rizzolatti, G., and Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience*, 27, 169-192, doi: 10.1146/annurev.neuro.27.070203.144230

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds

moving together. *TRENDS in Cognitive Sciences*, 10(2), 70-76, doi: 10.1016/j.tics.2005.12.009

Sperduti, M., Delaveau, P., Fossati, P., and Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: a brief review and meta-analysis. *Brain Structure and Function*, 216, 151-157, doi: 10.1007/s00429-010-0298-1

Sperduti, M., Pieron , M., Leboyer, M., and Zalla, T. (2013). Altered Pre-reflective Sense of Agency in Autism Spectrum Disorders as Revealed by Reduced Intentional Binding. *Journal of Autism and Developmental Disorders*, 44, 343-352, doi: 10.1007/s10803-013-1891-y

Strother, L., House, K.A., Obhi, S.S. (2010). Subjective agency and awareness of shared actions. *Conscious and Cognition*, 19(1), 12–20. doi: 10.1016/j.concog.2009.12.007

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, 17, 219-239, doi: 10.1016/j.concog.2007.03.010
Synofzik, M., Vosgerau, G., and Voss, M. (2013). The experience of agency: an interplay between prediction and postdiction. *Frontiers in Psychology*, 4(127), doi: 10.3389/fpsyg.2013.00127

Tsai, C-C., Kuo, W-J., Jing, J-T., Hung, D.L., and Tzeng, O.J.-L (2006). A common coding framework in self–other interaction: evidence from joint action task. *Experimental Brain Research*, 175, 353–362, doi: 10.1007/s00221-006-0557-9

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., and Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain,* 133, 3104-3112, doi:10.1093/brain/awq152

Wegner, D.M., and Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *The American Psychologist*, 54(7), 480-492, doi: 10.1037/0003-066X.54.7.480

Wohlschläger, A., Haggard, P., Gesierich, B., and Prinz, W. (2003). The Perceived Onset Time of Self- and Other-Generated Actions. *Psychological Science*, 14(6), 586-591, doi: 10.1046/j.0956-7976.2003.psci_1469.x

Wolpe, N., and Rowe, J.P. (2014). Beyond the "urge to move": objective measures for the study of agency in the post-Libet era. *Frontiers in Human Neuroscience*, 8(450), doi: 10.3389/fnhum.2014.00450

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA*, 310(20), 2191-2194. doi:10.1001/jama.2013.281053

Attribution of Agency to Automated Entities: Humanized versus Trained Systems

Michael Goldberg[1, 2], **Florian Koller**[2], **Niko Busch** [3, 4], **Elke van der Meer**[1, 2]

[1] Berlin School of Mind and Brain, Humboldt-Universität of Berlin, Luisenstraße 56, 10117 Berlin, Germany
[2] Institute of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany
[3] Institute of Psychology, University of Münster, Fliednerstraße 21, 48149 Münster, Germany
[4] Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Münster, Germany

Correspondence: Michael Goldberg, Institute of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany. E-mail: michael.goldberg@hu-berlin.de

Abstract

The sense of agency has been shown to take a new form when cooperating with another person in a joint task - an extended 'we' agentic identity. However, this agency attribution is overturned when cooperating with a non-human co-actor. We have tested two types of automated co-actors: a humanized and a trained computer. Additionally, feedback about who acted first (causing the outcome) was designed to match specific conditions and events. Our results showed that action binding was extended after a simulated training phase, but not with the humanized computer. In contrast, tones were shifted and extended for both co-actors. Similar to previous studies, feedback on who acted first had no influence on implicit agency. Our results suggest that a theory of mind might underlie the extent to which one feels in control when cooperating in a joint task, and that even direct "custom-made" feedback is ineffective with regard to pre-reflective agency.

Keywords: Sense of Agency; Intentional Binding Effect; Joint Action; Human-Computer Interaction; Anthropomorphism; Theory of Mind; Autonomous Systems;

1. Introduction

The sense of agency (SoA), most generally refers to the subjective feeling of control over one's bodily actions and the resulting sensory outcomes in the external environment. In recent years, this fundamental concept has attracted a growing amount of scientific attention and has been studied in a variety of different contexts (for a recent comprehensive review, see Haggard, 2017). A multitude of studies in the field have looked into the nature of the SoA, shedding light on its different levels, modulating factors and underlying neural mechanisms (see David et al., 2015 for a concise review about the current state of the research field; see Dewey & Knoblich, 2014 on the difference between implicit and explicit agency; see Sperduti et al., 2011 for a meta-analysis of underlying brain structures). Additionally, the SoA has also been investigated in a more applied framework, namely, clinical research, looking into psychopathologies characterized by disturbances of self-identity (among others, see Maeda et al, 2011 and Voss et al., 2010 for an investigation of schizophrenia and Sperduti et al., 2013 for autism).

More recently, two additional advancements have been made in the experimental research of the SoA. First, the applied framework was extended to the interface between the SoA and the field of human-computer-interaction (HCI). Among other central concepts, researchers have began to look at levels of automation (Berberian et al., 2012), input modalities (Coyle et al., 2012) and system feedback (Farrer et al., 2008), all of which have important implications to our feeling of control. This emerging research field is relevant to our everyday life, as interaction with automated systems is now integrated into almost all aspects of life. Findings about the perception, interaction as well as the general acceptance of technology inform both system designers as well as scientific human agency research (see Limerick et al., 2014 for more on the link between HCI and SoA; see Bagozzi, 2007 about technology acceptance). Second, the classical single person set-up was broadened to include multiple co-actors in a joint task. The step taken from the single to the multiple actor tasks opened-up a new range of questions. One of these questions is how we experience agency when cooperating with another person towards a common goal. Research conducted in this regard so far has uncovered a multitude of important insights relevant for co-action (see among others Dolk et al., 2014; Wohlschläger et al., 2003; Tsai et al., 2006). Among them, one interesting phenomenon is to be noted: when two people act together in a joint task, a new agentic identity is formed. The new agentic identity, a 'we' rather than an 'I', is generated insofar as the implicit, pre-reflective agency is extended beyond one's own actions and self-produced outcomes. That is to say, when a participant acts together with another co-actor, some sort of unified agency emerges and extends beyond the single actor (Obhi and Hall, 2011a).

Obhi and Hall (2011b) compared the attribution of agency to a human (confederate) with a computer co-actor. In order to test this comparison, the Libet-clock paradigm was employed and the intentional binding effect served as the implicit measure (for a detailed explanation of the Libet-clock paradigm and the intentional binding effect see Libet et al., 1983 and Wolpe & Rowe, 2014). Participants had to press a button, followed by a tone (i.e., operant condition), and estimate the times of actions and outcomes by noting the clock time of the event. These temporal estimations were compared to estimations on baseline conditions, where action is not followed by a tone and tone is presented without pressing the button. On each trial, either the

genuine participant or the co-actor (confederate/computer) acted first and caused the tone. One of the central findings showed a clear cut difference between the two conditions: when participants estimated the time of their actions and outcomes in the human-human condition, the action-effect interval was significantly shorter for operant compared to baseline conditions, pointing to an implicit SoA. Interestingly, binding effects were found for both the participant's own actions and outcomes, as well as when it is the co-actor's action that produced the outcome, implying an extended SoA over the co-actors actions and outcomes. However, when cooperating with a computer program instead of a human confederate, no binding effect (and so no implicit SoA) was found on any of the conditions. Not only was binding absent for action-effect intervals of the computer co-actor, but even the binding of self-produced actions and outcomes was under this context overturned. That is, neither extended nor self-experience of agency were found.

In a recent study (Goldberg et al., submitted), we embedded the joint task Libet-clock paradigm in a more natural, ecologically valid context, by employing a driving scenario. We used a driving simulator whereby participants were driving on a two-lane road side by side with another driver. The second driver was either a confederate, or a computer program, presented to participants as running the system of an autonomous car. The two-lane road was designed to converge at a certain point into a single lane. The goal of the joint task was to avoid an imminent crash at the convergence point by accelerating and overtaking the other driver in advance. Button presses and tones were in the new scenario converted to pedal presses (in order to accelerate) followed by tones indicating a successful avoidance of the crash. Participants performed the task twice with both the confederate and the computer, while unbeknown to them, the same behavior was employed by the program in both cases (that is, the confederate did not participate). The only difference between the two conditions was the belief manipulation about the partner with which the task was being performed. Unlike Obhi and Hall (2011b), we found that participants showed binding for the outcomes of both human and computer co-actors in the new meaningful context. Actions, however, were shifted only for the human co-actor, and not for the computer. Moreover, feedback about whose action took place first (in the form of the overtaking) showed to be ineffective on the implicit measure of agency. Our results suggested that the attribution of agency in a natural environment differs for actions and outcomes. While actions seemed to be closely linked to the agent producing them, and were more susceptible to the type of co-actor, outcomes were more affected by the meaningful cooperation of the joint task, and were therefore, shifted even in the human-computer condition. The results pointed to the fact the perception of action and outcome may, to some extent, rely on different underlying mechanisms and cues.

In the current follow-up study, we were interested to further test the difference between types of co-actors with regard to the attribution of agency. We were interested in finding out whether a sense of control can be regained by manipulating the computer co-actor in one of two ways: 1. Adding humanlike features to the computer's appearance; 2. Letting participants train the computer's system before performing the joint task. Specifically, we wanted to find out if either one of these manipulations would result in significant action binding (an effect not shown by previous studies). Both manipulations aim at closing the gap between computer and human co-actors. However, the two can be roughly categorized as being bottom-up

(first) and top-down (second) manipulations. Using humanlike external features might change the way in which the computer is perceived and thereby have an influence on the implicit measure. In contrast, a simulated training experience can possibly mold participants' attitudes on the cognitive level, thereby their temporal estimations. Additionally, we also investigated the influence of stronger, more direct feedback about the source of the action with regard to the implicit, pre-reflective level of agency. Since in both Obhi and Hall (2011b) and in our previous study, feedback did not influence the implicit level, we have strengthened and specified the feedback type with the intention of reaching a differentiation between the binding of self- compared to other- produced actions and outcomes. We hypothesized that it is the saliency of the feedback that had to be adapted for it to have an influence on both the explicit and implicit levels of agency.

Anthropomorphism refers to "imbuing the imagined or real behavior of nonhuman agents with humanlike characteristics, motivations, intentions, and emotions…" (Epley et al., 2007). In order to create an anthropomorphic version of the computer co-actor, we designed an avatar that identified as the autonomous car's system. In their study from 2014, Waytz and colleagues showed that an anthropomorphized autonomous car increased the level of participants' trust compared to a non-humanized car. The central features responsible for this effect were giving the car a name, gender and voice. We argue that the same psychological process that increases trust might underlie the inclination to unconsciously attribute agency to a non-human co-actor. Additionally, humanlike faces and bodies are known to promote the anthropomorphism of robots and other mechanical devices (Burgoon et al., 2000; DiSalvo et al., 2002). Therefore, to strengthen our manipulation, we also created a humanlike face and body that will go along with the other features, resulting in a more natural figure. We hypothesized that participants will be more ready to extend pre-reflective agency over an anthropomorphized computer's actions due to its greater external similarity to a human co-actor compared with the non-anthropomorphized computer (such as the one presented in our former study). We also predicted that when cooperating with the anthropomorphized car, actions and tones would be significantly shifted from baseline conditions, showing a full attribution of agency.

In order to implement our second manipulation, a second group of participants was asked to train the autonomous car's system prior to performing the joint task. In reality, no training took place and the computer's behavior was identical in all respects to the one presented in the avatar condition (excluding the avatar). The training phase included four stages, each dedicated to teaching the system basic driving skills relevant to its successful participation in the joint task (i.e., turning, accelerating, recognizing streets signs and overtaking). Participants were told that the system learns through their feedback, based on the elimination of wrong behavior. The joint task followed directly after the successful completion of the simulated training phase. We hypothesized that significant action and tone binding would result from a top-down process of inference, where participants form a better understanding of the "inner-workings" of the machine. In other words, we argue that the training phase might influence the ability to generate a humanlike theory of mind (ToM), whereby the mental distance between human and non-human co-actors is reduced. By theory of mind we refer to "our ability to explain and predict other people's behaviour by attributing to them independent mental states, such as beliefs and desires" (Gallagher & Frith, 2003). It is important to note that we do not appeal to the more

specialized conceptualization of ToM, which views a theory of mind as the ability to understand that someone else can hold a false belief (Hofmann et el., 2016). The latter definition, although widely accepted and tested with the help of false-belief tasks, is conceptually restricted and does not tap into the broader sense of what we wish to argue for. It is through a stronger sense of familiarity, that participants might regain a sense of control over their own as well as the other's actions and outcomes, in just the same way as they do when cooperating with a human co-actor (see Langer, 1975, for an experiment demonstrating the relation between stimulus familiarity and a sense of control).

To test the influence of feedback on the implicit pre-reflective SoA, the avatar was also employed to communicate with participants directly after each trial. Specifically, the avatar notified the participant about who acted first and was responsible for causing the tone (on blocks where action times were estimated), and who overtook the other driver (on blocks where tone times were estimated). Obhi and Hall (2011b) used patches of color to notify participants about the source of the action, and our previous study made use of the car overtaking as feedback. In comparison to that, we hypothesized that the additional emphasis and specification would result in a significant difference between temporal estimations of self- and other-produced actions and outcomes in the human-avatar condition. In the case that agency will be attributed to the computer co-actor, we predicted a stronger binding for self-produced actions and outcomes than for those produced by the computer. If no agency will be attributed to the computer, the self-agency would still be retained.

2. Materials and Methods

2.1 Participants

Forty undergraduate students from the Psychology Institute of the Humboldt-Universität zu Berlin (10 males, 30 females; mean age 28 years, SD=10.2, range 19-65) participated in the experiment. All participants were right-handed, had normal or corrected-to-normal vision and hearing, and held a valid driving license (mean of 7.5 years since its acquisition, SD=6.3, range 1.5-34). Participants gave written informed consent prior to the beginning of the experiment and were given course credit for their participation. The study was conducted according to the declaration of Helsinki (WMA, version October 2013).

2.2 Apparatus and Procedure

The experimental procedure followed our previous study (Goldberg et al., submitted) with a few additions and modifications described in the following sections. The experiment (both manipulations) was programmed and performed in Unity version 5.5.1 (Unity Technologies SF, US, 2009) and responses were registered using a wheel and pedal set (Logitech, Driving Force GT, E-X5C19). Visual information was displayed on a 27-in. DELL UltraSharp LED monitor, and sound was played through a connected headset.

2.2.1 General task and experimental set-up

Two identical set-ups were placed in the same room one next to the other (See Fig. 1). In order to avoid having any visual information about the co-actor's set-up, a dark curtain separated the two set-ups. Participants sat one meter away from the monitor and their gaze was directed at the center of the right visual field of the screen to simulate the position of a driver, driving on the right lane. We ensured that all participants could hold the wheel and reach the pedals comfortably, and that the sound was heard through the headphones while using the earplugs.

Figure 1: The experimental set-up. Two identical set-ups separated with a curtain were used for both types of co-actors. The participant sat next to the set-up on the right side. The other set-up remained empty and the server-like computer was turned on. The simulated training phase took place on the left set-up prior to the joint task.

Each trial began with both cars positioned one next to each other (participant's car on the right lane, the co-actor's car on the left lane) and a curvy road first had to be crossed. When reaching the straight part of the road, a street sign signaled the convergence of the two-lane road into a single lane (see Fig. 2). Once crossing the first street sign, a vertical bar (9 cm long) with 12 marks (ranging from 0 to 60, with steps of 5) appeared on the right side of the screen. Both the starting position of the bar (i.e., the amount to which it was filled) as well as its filling direction (i.e., up or down) were randomized each trial anew. The bar kept on filling up and down at a rate of 2.5 s in each direction, until reaching the second street sign (following Libet et al., 1983). Participants were asked to accelerate in order to overtake the other car and to avoid an imminent collision (i.e., the joint task). In order to achieve this, participants were instructed to spontaneously (without fixating on a specific point) press the gas pedal once in the area between the first and the second street signs. The pedal press was followed by a 250 ms interval after which a tone was played (1000 Hz for a duration of 100 ms). The tone signaled the successful avoidance of a collision and the overtaking then proceeded automatically (i.e., displayed on the screen). The overtaking was pseudo-randomized across the block (with equal amount of trials for each feedback type: 'self' or 'other', depending on who acted first) and indicated to the participant who pressed the gas pedal first and caused the tone. Each trial ended with the required estimation of either the time of the pedal press or the tone. For the

estimation, participants were presented with an identical empty bar and used a 24-position adjustment dial (placed on the wheel) to fill the bar to the estimated amount.

Figure 2: Operant trials. View from the participant's car during an operant trial (after crossing the first set of street signs). The vertical bar is seen on the right side and the second set of street signs can be spotted in the horizon.

2.2.2 Experimental design

Participants were assigned to one of two groups and completed the joint task with either the avatar co-actor or with the trained co-actor. The groups were matched and independent samples t-tests were performed to control for differences of gender (5 males and 15 females in each group), age (avatar: mean=28.5, SD=11.7; training: mean=27.6, SD=8.3; $t(38)=0.27$, $p=0.78$) and years of driving experience (avatar: mean=8.6, SD=7.9; training: mean=6.4, SD=3.9; $t(38)=1.07$, $p=0.28$). We devised a cover story that enriched and strengthened the belief manipulation of both groups. Participants were told that collaboration between the cognitive psychology group of the Humboldt Universität zu Berlin and the AutoNOMOS Labs group (Freie Universität Berlin) was formed in order to test a new autonomous car system. Moreover, to strengthen the belief manipulation, a bigger server-like computer was placed next to the second (left) set-up. The extra computer was turned on and a simulation graphics was displayed on its screen. Both avatar and training versions of the joint task experiment were identical in all respects. The unique features of each manipulation are described in section 2.2.3.

For each of the two groups, two types of conditions, baseline and operant, were presented in separate blocks. In the action baseline condition, participants were driving on a straight road towards a highway (see Fig. 3 left). Once a participant crossed the highway road sign, she was requested to accelerate by pressing the gas pedal once, in a spontaneous fashion (avoiding a pre-planned press). No tone followed. The participant had to estimate the time of her pedal press. In the tone baseline condition, the participant was driving in a mountainous landscape (see Fig. 3 right). A road sign signaled that the driver was about to enter a hazard zone and was requested not to press the gas pedal from that point on in order to decelerate. A tone,

signaling the end of the danger zone, was presented randomly in an interval of 2 to 6 seconds from crossing the road sign. The participant then estimated the time of the tone.
The operant, joint task conditions followed the scenario described in the previous subsection and were separated into different blocks, in which either the pedal press or the tone were estimated. As part of the cover story, participants were told that the tone would follow in a varying delay after the first of the two pedal presses. In reality, however, the tone always occurred 250 ms after the participant's pedal press.

Figure 3: Baseline trials. Before crossing the first set of street signs. **(Left)** Action baseline trials. **(Right)** Tone baseline trials.

In total, each participant completed two baseline blocks and two experimental blocks, and each block consisted of 15 trials. To control for learning, carry-over, fatigue and other order effects, four block sequences were created in advance and counterbalanced by assigning each participant to one of these four pre-determined sequences. At the beginning of the session, participants completed two free driving trials to get used to the simulator environment. Then, two normal trials followed, one for each estimation type (i.e., pedal press and tone), which were not included in the analysis.

After the experiment, participants were questioned in a short debriefing session about the purpose of the study, whether they noticed anything about the procedures or had any thoughts about the set-up. None of the participants showed suspicion with regard to the computer or the cover story's general reliability, including the feedback given on each trial.

2.2.3 Manipulations

The avatar manipulation included a female figure named Iris who presented herself to participants as the new autonomous car's system (see Figure 4 left). The central motivation for choosing a female figure was to control for gender difference between the current and the previous studies. In our previous study (Goldberg et al., submitted), participants cooperated with a female confederate. In order to be able to compare the findings of the two studies and make sensible inferences, we wished to keep the settings as close as possible. A set of pre-defined texts was interleaved before, between and after each block. Whenever the avatar talked to participants, her facial and bodily movements were synchronized with her speech, and her figure was displayed at the center of the screen. The avatar commented on the written instructions and reformulated them in her own words. For example, after the participant finished reading the instructions preceding an operant block (where the

action needed to be estimated), the avatar said: “Ah, we have to avoid a collision and then estimate the time we pressed the gas pedal”. Other times, the avatar would simply comment on the progress of the experiment. For example, after the third block the avatar said: “We’ve put the third block behind us. I think only one more block is left. Come on, you can pull through.”
Additionally, the avatar (head only, overlaid) also appeared towards the end of each trial on operant blocks (see Figure 4 right). As mentioned before, we used the avatar to specify and strengthen the feedback factor given to participants about who acted first and who caused the tone. For that purpose, the avatar was displayed directly after the tone sounded, and according to the type of estimated event (action or tone), a pre-recorded sentence was played. For example, if the participant had to estimate the time of the action, and pressed the gas pedal first, the avatar said: “You accelerated first” in order to emphasize who acted first. If, for example the participant had to estimate the time of the tone and the autonomous car acted first and overtook the participant’s car, the avatar said: “I overtook you” in order to emphasize who caused the outcome. The avatar’s face disappeared shortly before the estimation screen appeared.

Figure 4: Avatar Iris representing the autonomous car’s system. (Left) The avatar figure was presented before, after and in between blocks, and communicated with the participant. **(Right)** The figure that was displayed at the end of each trial and was used to indicate who caused the action first and thereby produced the tone (and the overtaking).

The training manipulation incorporated a simulated training phase that preceded the joint task. Participants were initially seated in front of the left set-up, and it was explained that they are required to train the system before driving alongside it. The system was described as learning from the participant’s feedback and through the elimination of false behavior. Participants watched the autonomous car driving on the left lane as it performed a set of pre-defined behaviors. The training was divided into four stages, each dedicated to train the system on a basic driving skill, which would have become relevant for its performance on the joint task later: turning, accelerating, recognizing a street sign and eventually merging into a one-lane road (See Figure 5). The required correct behavior was always described prior to each step. For each of the four stages we designed a set of wrong behaviors, deviating from the correct behavior to some degree, from completely wrong to almost correct. These were ordered to induce a sense of learning progress. For example, in the turning stage, the autonomous car would sometime turn to the wrong direction, not turn at all and drive out of the road or turn in the right direction but a bit too early, causing it to go off the

road. In between, a few correct trials appeared and each stage ended with two repeating correct trials. After each trial, participants had to decide whether the observed behavior was correct or incorrect. If incorrect, participants had to further specify which of the four options best describes the false behavior (presented in multiple choice form). At the successful completion of each step, a progress bar was presented, indicating the overall progress made by the system. After completing all four stages, participants moved to the set-up on the right side to do the joint task with the allegedly trained autonomous car.

Figure 5: Four stages of the training phase. Screenshots taken from the four different stages of the training phase performed prior to the joint task: turning **(top left)**, accelerating **(top right)**, street sign recognition **(bottom left)** and convergence **(bottom right)**.

2.3 Data Analysis

The experiment comprised a 2x2x2x2 multi-factorial, repeated measures design with the following factors: Co-Actor: Avatar, Training; Condition Type: Baseline, Operant; Estimated Event: Pedal press, Tone; and Feedback: Self, Other. In order to corroborate our hypotheses, we expected the results to show an interaction including condition type and estimated event (for interval analysis only condition type) together with either co-actor or feedback. To enable a comprehensive and detailed interpretation of the data, a dual analysis was performed following that of Obhi and Hall (2011b). Shifts of single events (i.e., action and outcome) and changes in the derived action-outcome interval across conditions were calculated. The method incorporates elements of both the traditional treatment of data in intentional binding experiments (such as in Moore & Haggard, 2008, which look at action and tone separately) and the more recent studies that go beyond the strict definition of the IB effect, uncovering a broader range of effects by also referring to the interval measure (Strother et al., 2010).

A measure for action temporal binding is calculated by subtracting the mean judgment error of actions in the action baseline condition from the mean judgment error of actions in operant conditions. The same holds for tone (outcome) binding

with the respective baseline and operant conditions. The size and direction of the temporal shift is taken to be the implicit quantitative measure of the extent to which the subject had the experience of agency. Our current design was based on the original Libet paradigm and followed closely the same parameters of the clock implemented in the vertical bar (i.e., rotation speed, rotation fashion and visual angle of the bar) and the estimation method, as well as all other decisive parameters including action-effect delay, tone characteristics etc. Since the Libet-clock paradigm is known to be a sensitive measure that can fluctuate on the trial level, extreme values were carefully excluded to avoid the distortion of results (for more on the factors influencing the time judgments using the rotating clock, see Pockett & Miller, 2007). The averaged means presented were therefore trimmed as follows: first, the standard deviations were calculated for each participant over all trials and blocks of each condition. Second, trials with values greater or smaller than three standard deviations from the mean of the specific condition were discarded together with values deviating by 500 ms from the time of the actual event. Finally, the trimmed means were calculated and averaged over all participants. The total number of rejected trials amounted to 2.6% of all trials.

3. Results

3.1 Single event analysis

Initially, judgment errors were calculated on the trial level (i.e., distance of the estimated time from the actual time of the event). Following, trial values were averaged for each condition type and estimated event, across all participants. The procedure was performed for both co-actors and feedback types. To test the hypotheses regarding the attribution of agency with both types of co-actors and investigate the differences in the temporal estimations of onset of single events, we ran a 2 (Co-Actor) by 2 (Condition Type) by 2 (Estimated Event) by 2 (Feedback) repeated measures ANOVA.

The analysis revealed several significant results: a main effect of condition type (operant conditions were significantly shifted from baseline conditions: $F(1,19)=3.56$, $p=.037$), a main effect of estimated event (stronger binding for tones than for actions: $F(1,19)=4.53$, $p=.047$), and an interaction between the two ($F(1,19)=44.7$, $p<.000$). A two-way interaction was found between condition type and feedback ($F(1,19)=5.66$, $p=.028$). Most relevant to our predictions, a two way interaction between co-actor and estimated event was found significant ($F((1,19)=5.55$, $p=0.029$). All other main effects and interactions were not significant.

To further investigate the interaction between co-actor and estimated event, paired samples t-tests were conducted (data were collapsed across feedback and condition type; Bonferroni corrected). A significant difference was found between action and tone binding when cooperating with the avatar, but not when cooperating with the trained computer (avatar: $t(19)=2.77$, $p=.012$; trained computer: $t(19)=.157$, $p=.877$). Post-hoc t-tests also showed that when participants performed the joint task with the avatar, temporal estimations of actions were not significantly shifted from the baseline (action forward shift: $t(19)=1.14$, $p=.133$), while those of tones were significantly shifted (tone backward shift: $t(19)=5.14$, $p<.000$). When performing with the trained system, both action (action forward shift: $t(19)=1.92$, $p=.035$) and tone

binding (tone backward shift: t(19)=2.95, p=.004) were significantly shifted. Figure 6 shows the mean temporal estimations of action and tone on baseline and operant conditions for the avatar (left) and trained computer (right) co-actors. In summary, tone, but not action binding, was found in the human-avatar condition, while both action and tone binding were significantly shifted in the human–trained computer condition. While a significant interaction was found between condition type and feedback, no difference between 'self' and 'other' temporal estimations was found on neither of the co-actors. The source of this interaction was revealed by running paired samples t-tests. Regardless of the type of co-actor and estimated event, the feedback of type 'self' showed stronger binding on operant compared to baseline condition (t(19)=2.96, p=.008).

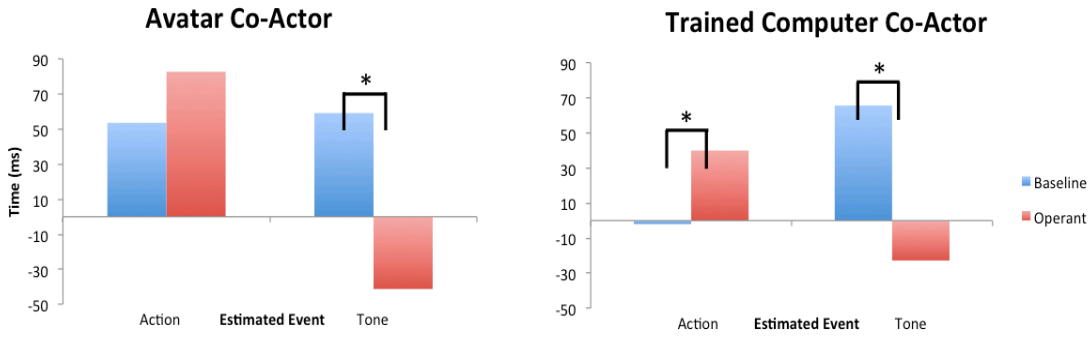

Figure 6: Single event analysis: mean temporal estimations. Mean temporal estimations of action and tone on baseline and operant conditions for the humanized **(Left)** and trained **(Right)** computer co-actors.

3.2 Interval analysis

To test agency attribution on the interval level, a second analysis was conducted. In order to investigate the differences between baseline and operant action-effect intervals, four intervals were first calculated. The intervals were matched by feedback (self/other) and condition type (baseline/operant): 1. Action-baseline-self – tone-baseline-self; 2. action-baseline-other – tone-baseline-other; 3. action-operant-self – tone-operant-self; 4. action-operant-other – tone-operant-other. Means were then calculated over all participants for each of the four intervals. Finally, a 2 (Co-Actor) by 2 (Condition Type) by 2 (Feedback) repeated measures ANOVA was performed. Since Estimated Event was already included in the calculation of the intervals on this analysis, it was not added to the ANOVA as a separate factor.

The analysis revealed a significant main effect of condition type (F(1,19)=44.7, p<.000). Action-effect intervals were significantly shorter on operant compared to baseline conditions, regardless of co-actor and feedback. That is, the binding effect was present on the interval level. In addition, it revealed a significant main effect of co-actor (F(1,19)=5.55, p=.029). While both types of co-actors showed a significant interval shortening (avatar: t(19)=4.18, p=.001; trained computer: t(19)=5.33, p<.000), the trained computer was shifted consistently more than the avatar. However, a post-hoc pairwise t-test showed no overall difference between the two (t(19)=.02, p=.983). All other main effects and interactions were not significant. These results partially reflected the results from the single event analysis, insofar as they supported a strong binding effect for both co-actors. Nevertheless, the fact that only tone but no

action binding were found on the avatar condition was not revealed in the interval analysis due to the strong magnitude of the tone binding on this condition. Figure 7 shows the calculated intervals on baseline and operant conditions, for the avatar and trained computer co-actors.

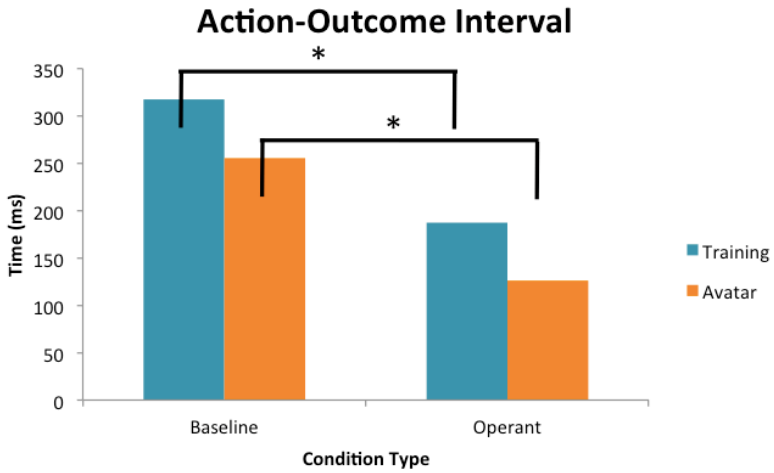

Figure 7: Interval analysis: mean derived intervals. Derived action-outcome intervals on baseline and operant conditions for both types of co-actors.

4. Discussion

We investigated the influence of two types of co-actors in a joint task on the attribution of agency (i.e., its extension) for actions and their outcomes. One group of participants performed a joint task with a humanized computer, which used a female avatar to represent the automated system. The second group started by training the computer's system in a simulated training phase, and subsequently performed the joint task with the trained system. Moreover, the impact of direct and specified task related feedback (i.e., the source of the action) on the implicit measure of a pre-reflective sense of agency was analyzed. For that purpose, we designed unique reactions played by the avatar, corresponding to each type of feedback, depending on the condition and event.
The manipulations were embedded within a driving scenario implemented by a 2-D driving simulator. Participants drove side by side with the computer co-actor and performed a joint task. Temporal estimations of self-produced actions and outcomes, as well as of actions and outcomes that were preceded by the co-actor, were collected. Data was analyzed on both the single event and on the action-outcome interval levels.

The study yielded the following main results: First, a significant forward action binding was found in the trained computer, but not in the humanized computer condition. This finding partially corroborates our predictions that participants would regain a sense of control when cooperating with the non-human co-actors through the manipulations. Second, a significant backward tone binding was found on both types of co-actors. This result replicates the findings from our previous experiment, where tone times were shifted for both human and computer co-actors, showing that outcomes rely less on the type of co-actor and more on the general context of the joint task. Third, action-effect intervals were significantly shorter in operant compared to baseline conditions on both the humanized and the trained computer conditions. A

further post-hoc t-test revealed no significant differences between the two interval shortenings. This result points to the fact that the tone binding was strong enough for the whole interval to be significantly shorter even on the humanized computer condition. Lastly, similar to the original study by Obhi and Hall (2011b), as well as to our previous study, we could not find any effect related to the type of feedback (self or other) given to participants on any of the conditions on the implicit level of pre-reflective agency. These results will now be discussed and interpreted.

The first part of the single event analysis, investigating action estimations, revealed a difference stemming from our two manipulations. Significant forward action binding was found when cooperating with the trained computer, but not with the humanized computer. The binding appeared regardless of whose action it was that took place first, meaning that self as well as extended SoA were experienced after the simulated training phase. These results mirror the pattern of results of the joint task with another human co-actor. In contrast to that, neither self nor extended action binding were found when acting alongside the humanized computer, showing no difference from the human-computer condition in our previous experiment (Goldberg et al., submitted). On a superficial level, the two manipulations, using an avatar in one group and a simulated training phase in the other group, are different: the one uses external cues to emphasize humanlike features in the non-human agent. The other enables participants to teach the system how it needs to act, increasing the sense of familiarity with its behavioral processes. However, we argue that the two tap into the same fundamental mechanism that reduces the gap between human and non-human partners. This is eventually a process of forming beliefs about, and understanding to some degree, the cognitive configuration of the co-actor, based on increased feeling of similarity to oneself (for more on the relation between joint action and ToM see a recent study by Humphreys and Bedford, 2011). If it is indeed the formation of a theory of mind that underlies both manipulations, the different results must be addressed. In other words, why is it that external cues don't work in the same way as the simulated training phase? An answer to that question might come from considering the complex notion of mind perception. In the past, the question of what entities have minds (maybe better restricted to cognition) was answered in a more or less straightforward way. For example, according to the Turing test, it is enough that one cannot tell the difference between human and non-human, for that entity to be considered as having some sort of mind (Saygin et al., 2000). Another view from the philosophy of mind is John Searle's strong artificial intelligence (AI) position: "The appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." (Searle, 1980). Although some may offer a clear-cut definition along these lines, we argue that perceiving an agent as having a mind is not an all-or-none type of process, but rather a matter of degrees. On the one hand, different aspects and faculties can be differentiated when trying to explain what it takes for an entity to have a mind. On the other hand, not only a multitude of aspects exist, but each of them can also admit to different degrees, considered subjectively by each person as crossing some implicit threshold or not. In their survey from 2007, Gray and colleagues studied the structure of mind perception and showed that minds are perceived along different dimensions. They have compared the judgments and estimations of different characters on different mental capacities. Two principal component factors were identified: experience and agency, each consisted of a series of different sub-constructs (e.g., the capacity to feel pain, having a personality, self-

control, planning, thought etc.). The fact that each character is rated differently on different mental capacities, shows that each aspect of agency is connected to different aspects of the agent and can be differentiated from the others. Therefore, it is possible that the training phase manipulation has tapped directly into beliefs about action generation, planning and control, not so much in the case of the avatar. While participants in the second group were given the chance to watch and correct the car's behavior, participants in the first group had only an indirect access to its decision-making process when presented with the avatar. The external similarity might therefore be related to other mental aspects of the agent (e.g., trust, empathy, ability to communicate etc.), but not to the pre-reflective perception of the computer's behavior.

Similar to our previous findings (Goldberg et al., submitted), backward tone binding was found with both computer co-actors. As this finding is now replicated, we are reassured in arguing that the difference in results from those of Obhi and Hall (2011b) stems from the newly designed context in which the joint task was performed. The nominal partnerships formed by merely watching the Libet clock were replaced with more meaningful settings. Participants were no longer simply involved in a symbolic task, but acted in order to achieve a common goal with overt consequences: when not acting in a timely manner, participants had to repeat the trial, and in the humanized computer condition, the avatar would show up and mention that to participants. When acting as required, an overtaking of either one of the two cars took place, at which point the avatar showed up once more. Moreover, the tones, having no conceptual relation to the button presses in the Libet paradigm, now played a significant role by carrying task relevant meaning. We argue that it was these contextual differences that account for the extended binding effect for both computer co-actors in the same way they did in our previous experiment.

In the interval analysis, we looked into the differences between action-effect intervals. The analysis revealed significantly shorter intervals in operant compared to baseline conditions for both computer co-actors. A further post-hoc t-test highlighted that the interval shortening was not significantly stronger on the trained computer compared to the humanized computer as could be expected from the single event analysis. Interestingly, similar to our previous study (Goldberg et al., submitted), tone binding was found to be stronger than action binding regardless of the type of co-actor (shown by the significant effect of estimated event), strengthening our hypothesis that action and tone rely on different types of external cues, one of which is the type of co-actor. In the current design, tone binding was strong enough to cause the interval shortening of the humanized computer condition to be almost as big as the one on the trained computer condition, where both action and tone binding contributed to the interval shortening. It is left a question for further investigation what could account for the greater magnitude of tone compared to action binding in all co-actors.

Finally, additional feedback delivered through the avatar's reactions after each trial had no significant effect on neither of the estimated events. Once again, while participants' belief was shaped by the explicit feedback (shown to be accepted as reliable), pre-reflective agency was unaffected. On the one hand, this finding contradicts our prediction that specified humanized feedback would be more effective than previous weaker types of feedback. On the other hand, this finding corroborates once again the distinction between the two levels of agency, that is, the 'feeling' and

the 'judgment' of agency. The implicit, pre-reflective level, as measured by the intentional binding effect, refers to an immediate non-mediated feeling of being the initiator of a certain bodily action. The explicit level of agency, or the judgment of agency, is a conscious reflection on the attribution of action and effect to either oneself or to another co-actor (Moore et al., 2012). Although not predicted, the feedback factor turned out to be significant in an interaction with condition type. This unexpected result is difficult to interpret, as it does not include the co-actor factor that was expected to explain the difference. Since feedback was strengthened only on the human-avatar condition (supported by its verbal iterations) and not on the trained computer condition, it is unclear what could the difference of 'self' feedback between baseline and operant conditions mean. One possibility is that participants were in general more attentive to feedback about their own actions, neglecting what might seem to be less relevant. This is in any case of no support to our predicted result. Further research is needed to analyze the impact of explicit feedback on the different conditions of the intentional binding paradigm in more detail.

In conclusion, the current study lent further support and extended our findings on the attribution of agency to non-human co-actors. By comparing two types of non-human co-actors in a joint task, we have shown that action binding was present and extended when cooperating with a trained computer but not with an externally humanlike system. The type of co-actor did not influence tone binding, as in both cases tone times were significantly shifted. Moreover, specified and emphasized feedback about the action source did not modulate the binding effect on the human-avatar condition as was expected, and proved the distinction between implicit and explicit agency to be unaffected. As discussed above, mind perception and the attribution of agency are highly complex notions. Although big steps are being made in both the theoretical as well as the experimental fields of agency research, some fundamental building blocks are still left to be uncovered. Progress in these fields is nowadays more important than ever before, as it is almost impossible to imagine humans running any daily routine without the help of automated systems.

5. Acknowledgments

We would like to thank the members of the Cognitive Psychology group at the Institute of Psychology of the Humboldt-Universität zu Berlin for their contribution to a fruitful discussion (and Benjamin Schlotter for programming). We would also like to thank Prof. Dr. Raúl Rojas, supervisor of the AutoNOMOS Labs group (Freie Universität Berlin), for permission to use the group's name and logo as part of our cover story. This work was supported by a doctoral grant of the Berlin School of Mind and Brain as part of the excellence initiative of the Deutsche Forschungsgemeinschaft (DFG) held by M.G. as well as by the BIGS² research school at Humboldt-Universität zu Berlin (F.K.).

6. References

Bagozzi, R.P. (2007). The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift. *Journal of the Association for Information Systems*, 8(4), Article 12. Available at: http://aisel.aisnet.org/jais/vol8/iss4/12

Berberian, B., Sarrazin, J-C., Le Blaye, P., and Haggard, P. (2012). Automation Technology and Sense of Control: A Window on Human Agency. *PLoS ONE*, 7(3): e34075, doi: 10.1371/journal.pone.0034075

Burgoon, J.K., Bonito, J.A., Bengtsson, B., Cederberg, C., Lundeberg, M., and Allspach, M. (2000). Interactivity in human-computer interaction: a study of credibility, understanding, and influence. *Computers in Human Behavior*, 16(6), 553-574, doi: 10.1016/S0747-5632(00)00029-7

Coyle, D., Moore, J., Kristensson., P.O., Fletcher, P.C., and Blackwell, A.F. (2012). I did that! Measuring Users' Experience of Agency in their own Actions. In *CHI, ACM Conference on Human Factors in Computing Systems*, (Austin, Texas, USA), 2025–2034

David, N., Obhi, S., and Moore, J.W. (2015). Editorial: Sense of agency: examining awareness of the acting self. *Frontiers in Human Neuroscience*, 9(310), doi: 10.3389/fnhum.2015.00310

Dewey, J.A., and Knoblich, G. (2014). Do Implicit and Explicit Measures of the Sense of Agency Measure the Same Thing? *PLoS ONE*, 9(10), doi: 10.1371/journal.pone.0110118

DiSalvo, C.F., Gemperle, F., Forlizzi, J., Kiesler, S. (2002). All Robots Are Not Created Equal: The Design and Perception of Humanoid Robot Heads. *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, London, England, 321-326, doi: 10.1145/778712.778756

Dolk, T., Hommel, B., Colzato, L.S., Schütz-Bosbach, S., Prinz, W., and Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. *Frontiers in Psychology*, 5(974), doi: 10.3389/fpsyg.2014.00974

Epley, N., Waytz, A., and Cacioppo, J.T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, 114(4), 864-886, doi: 10.1037/0033-295X.114.4.864

Farrer, C., Bouchereau, M., Jeannerod, M., and Franck , N. (2008). Effect of distorted visual feedback on the sense of agency. *Behaviopral Neurology,* 19, 53–57, doi: 10.1155/2008/425267

Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of 'theory of mind'. *TRENDS in Cognitive Sciences*, 7(2), 77-83, doi: 10.1016/S1364-6613(02)00025-6

Gray, H.M., Gray, K., and Wegner, D.M. (2007). Dimensions of Mind Perception. *Science*, 315, 619, doi: 10.1126/science.1134475

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18, 197-208, doi: 10.1038/nrn.2017.14

Hofmann, S., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L., Curtis, J., and Harris, P. L. (2016). Training children's theory-of-mind: A meta-

analysis of controlled studies. *Cognition*, *150*, 200–212, doi: 10.1016/j.cognition.2016.01.006

Humphreys, G.W., and Bedford, J. (2011). The relations between joint action and theory of mind: a neurophysiological analysis. *Experimental Brain Research*, 211, 357-369, doi: 10.1007/s00221-011-2643-x

Langer, E.J. (1975). The illusion of control. *Journal of Personality and Social Psychology*. 32(2), 311-328, doi: 10.1037/0022-3514.32.2.311

Libet, B., Gleason, C.A., Wright, E.W., and Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act.. *Brain*, 106, 623-642.

Limerick, H., Coyle1, D., and Moore, J.W. (2014). The experience of agency in human-computer interactions: a review. *Frontiers in Human Neuroscience*, 8(643), doi: 10.3389/fnhum.2014.00643

Maeda, T., Kato, M., Muramatsu, T., Iwashita, S., Mimura, M., and Kashima, H. (2011). Aberrant sense of agency in patients with schizophrenia: Forward and backward over-attribution of temporal causality during intentional action. *Psychiatry Research*, 198, 1-6, doi: 10.1016/j.psychres.2011.10.021

Moore, J., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, 17, 136–144, doi: 10.1016/j.concog.2006.12.004

Moore, J.W., Middleton, D., Haggard, P., and Fletcher, P.C. (2012). Exploring implicit and explicit aspects of sense of agency. *Consciousness and Cognition*, 21(4), 1748-1753, doi: 10.1016/j.concog.2012.10.005

Obhi, S.S., and Hall, P. (2011a). Sense of agency and intentional binding in joint action. *Experimental Brain Research*, 211, 655–662, doi: 10.1007/s00221-011-2675-2

Obhi, S.S., and Hall, P. (2011b). Sense of agency in joint action: influence of human and computer co-actors. *Experimental Brain Research*, 211, 663–670, doi: 10.1007/s00221-011-2662-7

Pockett, S., and Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition*, 16(2), 241-254, doi: 10.1016/j.concog.2006.09.002

Saygin, A.P., Cicekli, I., and Akman, V. (2001). Turing Test: 50 Years Later. *Minds and Machines*, 10, 463-518, doi:

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424, doi: 10.1017/S0140525X00005756

Sperduti, M., Delaveau, P., Fossati, P., and Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: a brief review and meta-analysis.

Sperduti, M., Pieron , M., Leboyer, M., and Zalla, T. (2013). Altered Pre-reflective Sense of Agency in Autism Spectrum Disorders as Revealed by Reduced Intentional Binding. *Journal of Autism and Developmental Disorders*, 44, 343-352, doi: 10.1007/s10803-013-1891-y

Strother, L., House, K.A., Obhi, S.S. (2010). Subjective agency and awareness of shared actions. *Conscious and Cognition*, 19(1), 12–20. doi: 10.1016/j.concog.2009.12.007

Tsai, C-C., Kuo, W-J., Jing, J-T., Hung, D.L., and Tzeng, O.J.-L (2006). A common coding framework in self–other interaction: evidence from joint action task. *Experimental Brain Research*, 175, 353–362, doi: 10.1007/s00221-006-0557-9

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., and Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain,* 133, 3104-3112, doi: 10.1093/brain/awq152

Waytz, A., Heafner, J.and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113-117. doi: 10.1016/j.jesp.2014.01.005

Wohlschläger, A., Haggard, P., Gesierich, B., and Prinz, W. (2003). The Perceived Onset Time of Self- and Other-Generated Actions. *Psychological Science*, 14(6), 586-591, doi: 10.1046/j.0956-7976.2003.psci_1469.x

Wolpe, N., and Rowe, J.P. (2014). Beyond the "urge to move": objective measures for the study of agency in the post-Libet era. *Frontiers in Human Neuroscience*, 8(450), doi: 10.3389/fnhum.2014.00450

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA*, 310(20), 2191-2194. doi: 10.1001/jama.2013.281053